Image Sampling Myths

Craig Stark

In the previous entry in the series, I went over a number of aspects of image sampling. In it, I argued that running at rates much more than a half (or perhaps a third) the FWHM of your skies wasn't going to buy you anything more in terms of resolution in your shots. At the same time, I argued that you're going to get a hit in your SNR. I purposely didn't come down hard on the "f-ratio myth". I danced around a few issues there and left things pretty broad by saying that "f-ratio doesn't rule the day and account for everything, but it also isn't entirely irrelevant." A great thing about Cloudy Nights is that there are forums there and readers can call you on it. In particular, I have to thank three individuals for pushing me on this: Sergi "xatamec", Frank "freestar8n", and Mitch Cox. I'd been thinking of developing this a bit more but these three get the credit for giving me a rather large shove. Thanks gents!

So, in this entry, we'll cover three things that have seemed to become dogma in our hobby, each of which has some problems that lead me to reject the simple mantras. The first is on Nyquist and as it pertains to image sampling and the second is on f-ratios. Both expand on Part 4, so if you've not read that back, you may find it useful to do so before continuing. The third is really an extension of the second and covers how much of this debate is driven by film vs. CCD.

Myth 1: The Nyquist Theorem doesn't apply to our images

The first thing on the docket is the notion that the Nyquist Theorem (and its sampling limit) applies only to things like one-dimensional audio streams and does not apply to things like two-dimensional images. This is pure bunk. If you have an two-dimensional, three-dimensional, or even N-dimensional analog source of information and wish to digitize it, the Nyquist Theorem applies. It is a theorem about information. It and Fourier analyses don't give a whit about the number of dimensions they are working on.

That said, the Nyquist Theorem has a number of stipulations. It works perfectly when you adhere to these stipulations, but when we don't, it can break down a bit and we won't get a perfect reconstruction of our data. We'll come back to this a bit at the end. But to begin with, what is the Nyquist Theorem and what is Fourier analysis?

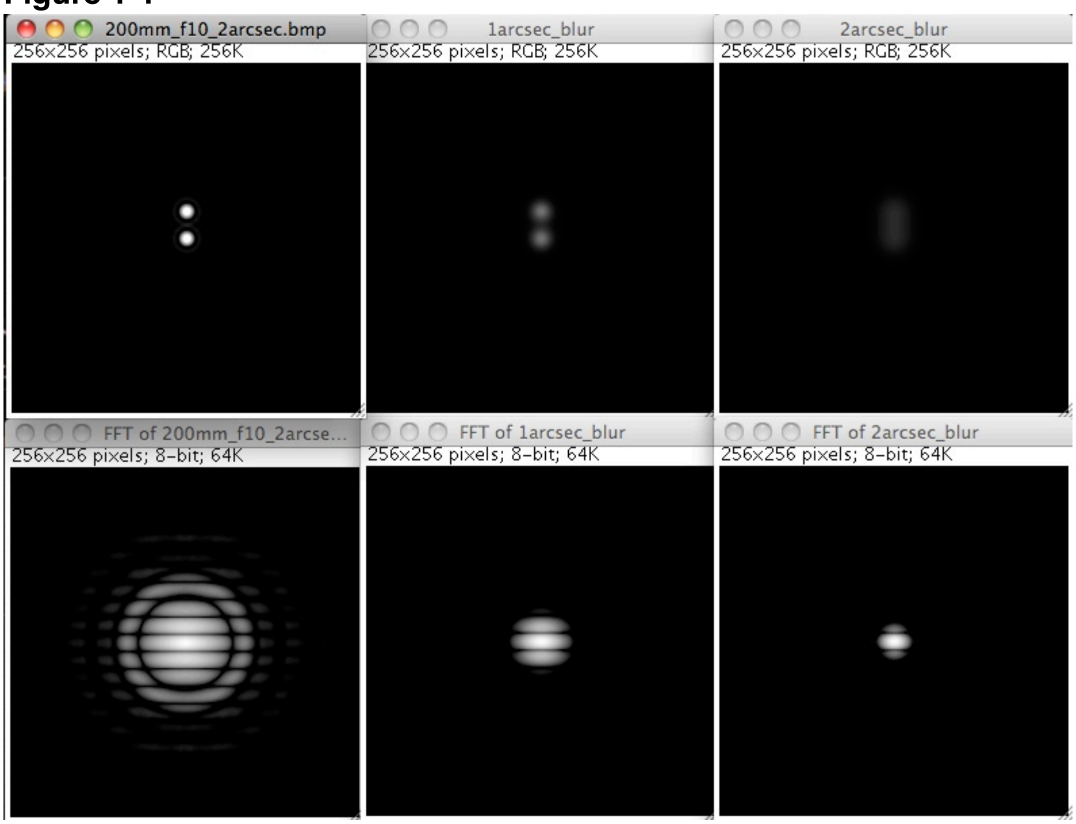
To start with the latter, in its commonly-applied form, Fourier analysis breaks a stream of information into a set of component frequencies (and complex phases). This transformation is invertible such that you can reverse it and re-create the original stream. The stream can vary along time (as we have in an audio waveform) or in space (as we move across an image in each direction). Once in the frequency (and complex phase) domain, we can look for and potentially filter out certain components. (This is similar on the surface of things to what Registax does in letting you filter or enhance different wavelets.) So, we can use it to detect things (like RF noise in your camera - see Part 3: Measuring your Camera), to enhance frequencies (e.g., enhance relatively high frequencies to sharpen an image), to reduce frequencies, to compress an image, etc. It's really an amazingly useful concept and our modern technological world couldn't exist without it.

Fourier analysis says that we can break down any image into its component frequencies and then invert this transformation to re-create the image. I demo'ed it in Part 3 with an image of my son. The question then comes up: if we are going to digitize this, how well do we need to sample it? What is the relationship between frequency in the image and sampling rate? This is where Nyquist comes in.

The Nyquist Theorem says that in order to reproduce frequency X, you must sample the stream at a rate of at least two times X. So, for a 20 kHz tone (most readers can't hear this high anymore), you need to sample it at 40 kHz. Now, in the conversion from analog streams (e.g., music coming from a trumpet) to digital, we have to worry about aliasing. Frequencies in the stream that are higher than the sampling rate get *aliased* back in as lower frequencies (think of beat-frequencies if you know what they are to give a basic idea of the process. In practice, it comes through as noise.) So, we must filter them out and it's the construction of these filters that leads to CDs having sampling rates of 44.1 kHz rather than 40 kHz. Imperfections in making the filters lead to the desire to have a bit of breathing room here. Thus, in digital systems, you'll often find things running a bit higher rates than Nyquist would require. This, also, is where we have a bit of an issue with our images. Most cameras don't have a lowpass spatial filter to block the higher frequencies that will alias into the image as noise.

Aside: You may be thinking -- Hey, that Nyquist thing says that the frequencies are sine waves. What if I had a 20 kHz triangle wave or square wave? It'd reproduce that as a 20 kHz sine wave after all that Fourier stuff and conversion. Well, yes, it would. Actually, coming out of the DAC would be a 20 kHz square wave. But what makes that 20 kHz wave square are very high frequency components that turn the sine wave into a square (or near-square - perfect square waves don't exist). Remove those very, very high frequency components and you get back the 20 kHz sine wave. So, that 20 kHz non-sine wave has frequencies higher than 20 kHz. If we want to accurately reproduce them, we need our sampling rate to be a lot higher than 40 kHz.

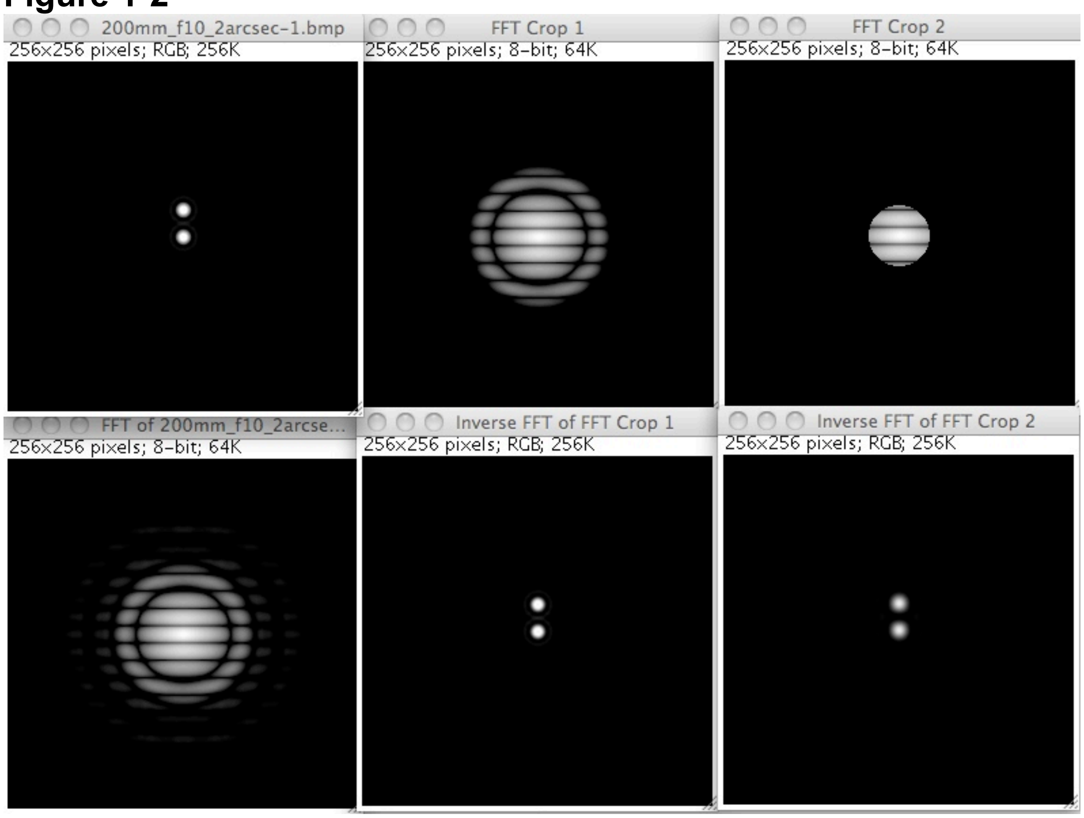
Figure 1-1



That concern aside, we can demo this ourselves to get a feeling for how this works using readily available software (Aberrator and ImageJ). First, I used Aberrator to make some images. The ones I'm working with here are of a double star, 2" apart stars in an 8" scope (optically perfect, upper-left of Figure 1-1). ***So, this is a pretty tight double for us to be resolving in our DSO images.*** Then, I brought these into ImageJ (see Part 3 for details on ImageJ) and made two copies of the image. I did a 1" FWHM blur on one and a 2" FWHM blur on the other. Recall that ImageJ uses sigma, not FWHM to specify the blur size (as does Photoshop). So, the blur sizes were 4.25 and 8.5 pixels (the image scale is 0.1" / pixel, so 2" FWHM is 20 pixels, and divided by 2.35 to convert to sigma and you get 8.5). I then did an FFT on both of them and you see the resultant loss of high frequency detail in the FFT of the blurred image. Remember, the center of the FFT image is the lowest spatial frequency and moving out from the center shows the energy at higher spatial frequencies. So, if you have 1" or 2" worth of seeing, this is what you get.

Note how you can see the image soften with blurring as we move across the top row and how the FFTs show less and less energy at higher frequencies. Note in particular how much high frequency information has been lost with only 1" FWHM of blur added to the image (the FFT is now very compacted in the center). To help us see how FFTs and frequencies work, I next took that first one (the unblurred image from the upper left) and cropped the FFT with a circular mask to cut out about half the frequency space and did an inverse FFT of it. The result is in Figure 1-2 (top row is the "input", bottom row is the "output").

Figure 1-2



You can see we can crop out a lot of the frequency domain (the high frequencies) and do very little to the image. The inverse FFT of the upper-middle (shown in the lower-middle) looks very similar to the original. Those outer rings have very little energy even in this perfect image and you can't really tell upper-left from lower-middle. Cut out a lot and you can (a harsher crop of the Fourier data is shown on the right side), but the point is we certainly don't need 0.1"/pixel resolution even here on this perfect image. Not only does the FFT show the whole frequency space isn't being used (lower-left), but a good bit of that is even carrying very little information. Heck, if we cut things very hard (as on the right) we still get something reasonable.

Notice, now that the very hard one has a hard circular crop in the frequency domain about where our pure image, blurred by 1" FWHM had all of its energy (Figure X1, lower-middle). The radius of that is about 25 pixels (or about 1/10th the image). What that means is that there is no energy at a frequency higher than this. Given the energy content, 1" is all the resolution we have, so we can sample this at 0.5" and not lose anything at all. In Figure 1-3, I took that image, rescaled it in ImageJ down to 48x48 pixels (a 5.3x reduction using bilinear interpolation, putting us at 0.53"/pixel) and then blew it back up with bicubic resampling. This shows that as long as we sample the image appropriately, nothing need be lost.

Figure 1-3

Figure 1-2 showed that even if we undersample our image rather dramatically (we cut out a lot of the data in the Fourier domain by that circular crop), we can still have a good amount of detail. In any case, as we get into real world situations with 1", 2", 3" or sometimes more worth of blur, running at rates higher than the Nyquist rate isn't going to buy you anything, even on highly magnified stars. Keep in mind, these two stars are only 2" apart. We're looking at the most detailed kinds of things you're looking to resolve and this simulation here is with only 1" FWHM of seeing. My skies are routinely a lot worse than this. Figure out the limiting factor in your seeing (be it skies, tracking, or driven by diffraction given the aperture of your scope), halve that and don't consider it a good thing to go beyond this. Even if you're on the other side of this and aren't "critically sampling" (or oversampling) the image, the high frequency component you've lost may not have that huge an impact on the overall image. Look again at the middle column in Figure X2. There, I cropped off some of the high frequency information before doing the inverse FFT. Can you really tell the difference between the upper-left and the lower-middle panel? After stacking a bunch of pictures would any difference you see now still remain?

So, to sum up, the Nyquist Theorem says that if we follow all the rules, we need only sample an image at a rate that is twice the highest spatial frequency we want to be able to faithfully reconstruct. We'll bend the rules a bit at times and we have things like how f-ratio affects star shapes, how CCDs are imperfect, etc. that will place other limits on our performance (ask Frank freestar8 for more details on this and be prepared to take notes!). So, if the utmost fidelity in spatial resolution is the goal, going a bit beyond the 2x rate is a good idea. For many of us, the amount of information in those very high frequencies may not be enough to justify the higher sampling rate and we do need to keep in mind that our skies are placing strong limits on what most of us can achieve. We then come back to the suggestion that if you're going for the most spatial detail possible, running at half or perhaps a third of your seeing is really all you should aim for. If you're looking for something more general-purpose (and if you don't want a hit in SNR), you should aim for something wider as we'll see in Myth 2.

Myth 2: For SNR, the f-ratio doesn't matter or matters only in extreme or marginal cases

This is a far more contentious issue. So, with flame retardant suit donned, I will continue, ready to be considered disparagingly in some circles. Much has been said about the role of the f-ratio in our signal-to-noise and much of it has not been clear. As noted at the outset, I've not been entirely clear on it myself. This is my attempt to set the record straight so far as I see it. One thing to note at the outset is the way I see it is biased by the kinds of images I want to take. I'm not into photometry and pictures of double-stars don't get me going either. I'm a DSO guy who treats astrophotography as a technical art. I have nothing against science (I'm a scientist by day), but my goal here is to make nice images of DSOs. These are inherently extended objects, so I pay particular attention to how well we can image them. This places more weight on one side of a trade-off - the trade-off I brought up in the last installment. But, as with the last installment, the key here is that a trade-off exists.

So, if you've got the same aperture of scope, does the SNR of the image get affected by what f-ratio you're running at?

Pixel-based vs. Object-based SNR

To begin with, we have to come to grips yet one more time with the concept of SNR. Specifically, we need to be clear about what "kind" of SNR we're talking about. We have to do this because the popular "CCD f-ratio Myth" page by Stan Moore has a different take on what kind of SNR we should be talking about. To understand where my take splits from Stan's and to understand why I've said in the other articles that none of the text contradicts his work, we need to delve into the two definitions of SNR. In all this, we'll hit again something akin to one of us looking at the glass as half full and the other as it half empty. Since I took the half-empty one last time, I'll give that honor to him this time.

First off, we should point out where there is clear agreement:

- Stan says, "There is an actual relationship between S/N and f-ratio, but it is not the simple characterization of the f-ratio myth." *Agreed. One f-stop will not double the SNR as many would have you believe.*
- He also says, “Information about an astronomical object (star, galaxy, nebula, features of galaxy or nebula, etc.) is contained in the light that falls onto Earth. That light consists of a certain number of photons per second per square meter of earth’s surface. The quality of information from an object depends on how many photons are captured and measured by the instrument.” *Agreed again. But, we’re starting to see the point of divergence here as he is talking about “information”. We should also note that the photons need to arrive at a portion of the image plane that you’ve got covered by your sensor.*

For Stan, the “true SNR” is “object SNR” and this “refers to the actual information content of the image”. To understand this, you must think about that term “information”. Think of it as how much data are really in the image or how many bits it would take to compress the image without losing anything (have a look at Shannon Information Theory if you want to read up on this). For example, a dark background with a bit of a galaxy core showing through has less information than a dark background with a nice bright galaxy showing through. If you capture more photons (e.g., with a bigger aperture) you will have more information, all else being equal. So far, so good.

Likewise, if you have the same galaxy image overall such that a thumbnail version of two images would look the same but one has a lot of detail in the arms and the other is blurred, the more detailed one has more information than the more blurry one. The more information in your image, the better. Here, again, I will agree in principle.

The problem is, by using the term “information”, we’ve muddied the water a bit. It’s held up as a Holy Grail so to speak or at least a lot more interesting than the more mundane pixel-based SNR. While it can be treated purely objectively (and we can calculate exactly how many bits of information are in an image), for us, there is often a subjective aspect to it. There is the real concept of useful information vs. useless information. What if we don’t care so much about spatial resolution (spatial information) and are willing to sacrifice some of those details by undersampling a bit so that we can get a much wider field of view? We’ve gained a good bit of information (many more stars and galaxies in the image, for example) but lost some in the process (detail in the arms of said galaxies). Is one better than the other? Should I really be working to maximize a raw measure of information here? I don’t think so. It’s a useful way to think about things, but I don’t think we must diligently work to maximize this value. Heck, I’m a sucker for wide-field shots, having clearly lost some spatial resolution (gaining and losing information here in the process). I’m not sure I care where Shannon would come down on this. To me, the wider field of view’s added information more than offset what I lost in terms of resolution. Were I interested in planetary nebulae, I’d sing a different tune, however.

I also don’t think that it’s the best way to think of SNR as it pertains to making a nice, clean image of DSOs. I agree entirely that the bigger the aperture of your scope the more photons you will capture from a given target, assuming that target fits on your CCD chip. That’s a given. But, each CCD well is largely independent (and CMOS sensors are even more so). Each pixel’s job is to estimate to the best of its ability how many photons hit it and it doesn’t care a) whether the photons are from a DSO, the skyglow, a star, or from heat and b) what the CCD well next to it is doing. It’s not like all the pixels that are part of the DSO all get together to share notes and secret handshakes. Each is just a detector. This detector is described by that simple pixel SNR equation.

Now, when we make an image out of these pixels our eyes and visual systems do impose a relationship among nearby pixels. We form lines and edges and we're very good at picking them out amidst noise (read up on "Gestalt psychology" for some fun demos of this). But, as we're doing our various DDPs, curves, and levels, we're just taking each pixel's value for what it is and shifting it according to a transfer function (e.g., an S-shaped gamma, an arbitrary curve, or a power function). If the estimate of the true intensity value for each pixel is closer to the actual truth (i.e., if there is less noise), we can stretch things more before the image breaks down. That is, if the pixel SNR is higher, we can stretch it more before the image looks noisy.

Thus, it is the pixel-based SNR that I worry about. This is the SNR we have defined in the other entries in the series and it is this SNR that you will find in typical discussions of the SNR in images and in our CCD cameras. We talk about the estimate of the number of photons captured by that pixel for that pixel's area of sky (not the whole detector's area of sky or the entire DSO's area of sky) and the variance in that estimate. As someone who is interested in imaging extended objects and picking them out as cleanly as possible from the background, this is a very important kind of SNR. I care about this first and spatial resolution second. Detail in a galaxy arm cannot be had unless you can record that galaxy arm in the first place. (Note, when I have talked about "Target SNR" in other articles and here, I'm talking about the SNR in that pixel for the photons from the target. I do this to make sure that skyglow isn't part of our "signal". It's not talking about SNR for the entire target, but rather is the "Target Pixel SNR").

So, Stan and I are talking about two fundamentally different concepts as we use very different definitions of SNR. This is why I have given a wide berth to his coverage and said that none of what I have gone over has really contradicted him. We've been talking about two different things. That said, not only this has led to some real confusion, but it's not been entirely accurate. For, even when you take real "information" into account (i.e., you have critically sampled the image so you have not lost any spatial resolution - see Myth 1), f-ratio still has an effect and it's not a non-trivial one. The effect is going to come down to not only a role for the read noise in the camera, but also a role for the shot noise from the target that hasn't always been considered.

The Role of f-Ratio on Pixel-SNR: The Equations

Let's go back to the full pixel SNR equation (as always, having the "signal" be the photons only coming from the target and not the photons coming from the skyglow and dark current as we want a pretty picture of the target). We'll do this to drive home the point that in terms of pixel-SNR, f-ratio is going to matter. We have:

$$TargetPixSNR = \frac{TargetSignal}{\sqrt{TargetSignal + SkyglowSignal + DarkSignal + ReadNoise^2}}$$

For now, let's pretend we have a perfect camera with no dark current and no read noise. We thus just have our target's signal and the skyglow's signal:

$$TargetPixSNR = \frac{TargetSignal}{\sqrt{TargetSignal + SkyglowSignal}}$$

Now, let's consider what happens if we keep the aperture constant and we change the f-ratio by one stop. Say, we went from a 100 mm f/5.6 scope to a 100 mm f/4 scope. We've shifted from 560 mm to 400 mm of focal length. Were your SLR hooked up to these scopes, it would compensate by halving the exposure duration. Why? Each pixel (and the whole sensor) is getting twice as much light (assuming you're shooting a flat frame or something) as it's covering twice as much sky. Run the math if you like and you'll see that each pixel is covering sky 1.414 times as wide and 1.414 times as high (sqrt(2)). You can also think of this as making each pixel physically bigger by a factor of 1.414 in each direction.

Pretend for the moment that this pixel is aimed at a relatively smooth part of some nebula. So, expanding its FOV hasn't made it now in some brighter or darker patch. It's the same basic stuff, only more of it. What happens to the SNR? Let's call this TargetPixSNR′ (the ′ to separate it from the first one we calculated). We get:

$$TargetPixSNR' = \frac{2 \times TargetSignal}{\sqrt{2 \times TargetSignal + 2 \times SkyglowSignal}}$$

Simplify this a touch and we get:

$$TargetPixSNR' = 1.414 \times TargetPixSNR$$

Thus, by having a one-stop change in the f-ratio and keeping the aperture constant, we have boosted the SNR here by a factor of 1.414. We boosted the signal by a factor of 2 (which is why your SLR will halve the shutter speed), but the SNR went up by only 1.414. That said, it didn't go up by something like 1.0001. No, it went up by 41%. That's not chump change.

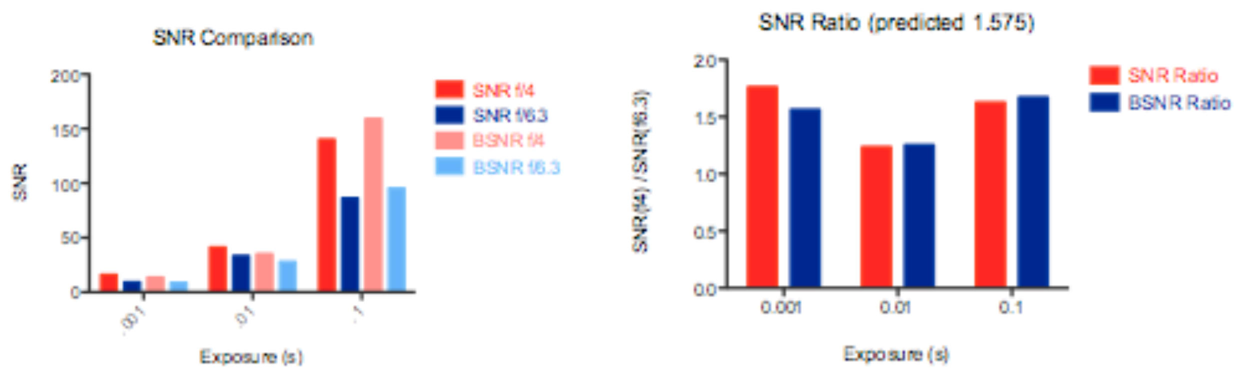
Note, the 1.414 here comes from it being the ratio of the f-ratios (or the ratio of the focal lengths since the aperture here is constant). It's also the square root of the boost in the amount of light. Here, we doubled the amount of light (by doubling the area of sky) and sqrt(2) = 1.414. If we'd cut the focal length in half down to 280 mm, we'd have quadrupled the amount of light and boosted the SNR by a factor of 2 since sqrt(4) = 2. Again, we can get to this by the square root of the boost in light or by just the ratio of the focal lengths or f-ratios (560 / 280 = 2 and 5.6 / 2.8 = 2).

The Role of f-Ratio on Pixel-SNR: An Insanely Simple Application
Some readers don't like it when things are done in "just math". Let's take an actual image but one we can control here. We want something with an actual scope and an actual camera, but we want the target to not be affected by seeing, transparency or actual contrast variations (we're looking at an even part of this nebula such that our pixel's coverage is still in the same basic intensity range despite covering more sky). One such target would be a flat frame. We can use this as a proof of concept for extended objects (again, not stars and yes, we are ignoring the obvious fact that spatial information will lost here).

Here, I took my Borg 101 ED run at its native 630 mm focal length (f/6.3) and run at its reduced focal length of 400 mm (f/4) and shot my EL flat panel using my QSI 540. The scope was focused at infinity and 20 frames were taken at 0.001, 0.01, and 0.1 seconds using each optical configuration. Images were bias corrected using a large stack of bias frames. For the f/6.3 rig, the mean intensity in the center of the image ranged from 154-13.8k ADU and for the f/4 rig, the mean intensity in the center of the image ranged from 357-31.8k ADU after this processing.

Now, according to the math above, the boost in the SNR here should be 630 / 400 or 1.575. I calculated the SNR in my test images two ways. The quick and dirty way many would do it is to just take the mean intensity in a locally-flat area (here, a 10x10 area in the middle of the image) and divide it by the standard deviation in that area. This is a good proxy for the SNR, but it does assume the image is perfectly flat and that the sensor (post bias correction) is perfectly flat. To clean this up a bit, I did it the more exacting way as well, calculating the mean value and standard deviation for a given pixel across the 20 images I took. The former is the "SNR" and the latter the "BSNR" (better SNR) in the graph here in Figure 2-1.

Figure 2-1



On the left, you can see the SNR for the f/4 configuration is higher regardless of the image intensity or the measure of SNR. On the right, we see how much higher it is. With a perfect camera, no extra light loss associated with the reducer, no non-linearities in the detector, etc., this should come out to a 1.575x improvement for the f/4 condition over the f/6.3 condition. The mean of the SNR bars here on the right is 1.54 and the mean of the BSNR bars on the right is 1.5. I'd say the math held up pretty darn well.

So, in this semi-real-world test, a shift in the f-ratio, holding aperture constant boosted the SNR of an extended object by the predicted amount. Reducing the f-ratio by a factor of 1.575 boosted the SNR by this amount as well. *Thus, f-ratio is clearly having an effect on SNR and it's not something to be swept under the rug.*

Getting Back to Information

At this point, those who uphold the notion that f-ratio's effects are a myth may be rather incensed as what I've done here is to present a case with very little real information in it. A perfect image of a flat has almost no information in it. You could represent all 4 million pixels in my camera with a single value, say 1327 if the optics were perfect and there were no noise. A single value (repeated across all pixels) is very little information. ***The point of that exercise, however, was to show that when we're not talking about gaining or losing some spatial detail, the f-ratio surely does matter in our ability to estimate the intensity of that spot of the nebula.*** Move over some number of pixels where the nebula is perhaps darker and we'll have a lower value. Our SNR in that pixel will be higher with a lower f-ratio. Thus, when we stretch the image to better reveal the difference between these areas, we'll have a lower-noise image and be able to show the contrast between the two areas better.

What happens when we have actual spatial information in here we're worried about losing? If we turn back to Myth 1 here, we note that oversampling an image much beyond what Nyquist would indicate isn't letting you capture any more information. You have a certain amount of information in the what is passing through the cover-glass on your CCD and that amount is dictated by the inherent blur in the image. That blur is dictated by the effects of aperture (larger apertures reduce the size of the Airy disk), seeing, tracking, etc. At this point, how it is sampled and what the f-ratio are doesn't matter.

But, we must take that information as it is passing through the cover-glass on the CCD and we must sample it and record it using real detectors. Let's pick apart the two aspects of information here: spatial information and intensity information.

For spatial information, we must consider the sampling rate. If we are at or above the Nyquist rate, we are not losing spatial information. If we are below this rate, we are losing spatial information. How much we are losing depends on how badly we are undersampling the image. Depending on how we have gotten to this undersampling point, we may or may not have gained information. If we have a sensor of the same physical size and have just increased the size of the pixels (e.g, by binning), we have lost spatial information (again, if we are below the Nyquist rate). If we have kept the same sensor and have done something like put a focal reducer in place, we have lost spatial resolution but we have gained field of view (FOV). Clearly, those extra stars, galaxies, nebulae, etc. are information. Whether it is useful information or not is in the eye of the beholder. But, if your arcseconds per pixel is more than roughly half your seeing (or whatever else is limiting your resolution), you're certainly starting to lose some spatial resolution. Whether that trade-off you've made is worth it or not (e.g., by exchanging some spatial resolution for greater FOV), again is in the eye of the beholder. This is technical art and when we consider changes in FOV, the technical bits like consideration of information-content break down.

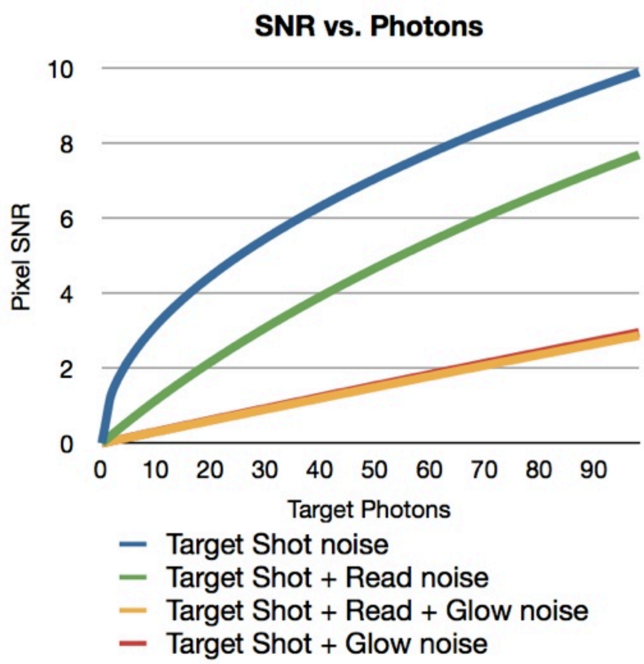
For intensity information, as shown above, we must consider the fact that these are real-world detectors of photons and that photons behave according to a Poisson process (see Part 1). We have read noise and we have shot noise. Ignoring read noise, if you pack N times as many photons onto a CCD well, your SNR will be sqrt(N) higher as a result of this Poisson process. How do you get more photons to pack into a CCD well? Alter the f-ratio. The aperture will determines the total number of photons from the sky that are collected. The focal length determines the size of the image and therefore how they are spread over the sensor. The f-ratio, being the ratio of these two, determines how many photons are packed into each square millimeter (or micron, or what have you).

If we go back to Part 4 and look at the equation for the image scale, we see that the number of arcseconds per pixel is based on the focal length (206.265 * PixelSize / FocalLength). So, if we keep the focal length constant, we keep each pixel covering the same amount of sky. If we want to get more photons hitting the detector from this same amount of sky, how do we do it? We increase the aperture. Aperture wins? Not so fast, because in so doing, we dropped the f-ratio. Same focal length, but larger aperture results in a lower f-ratio. Take that increased aperture's extra photons and spread them out more now so that you have the same number of photons hitting each CCD well as you had with the smaller scope. To do that, we've had to boost the focal length, making a larger image on our CCD and having each pixel cover a smaller area of the sky. We've boosted the f-ratio back and, in fact, we've boosted it to exactly the level we were at before. The f-ratio is tracking the number of photons for an extended target perfectly. That's what it's designed to do. There's nothing magical about film or CCDs here. It's just simple geometry.

Can we claim that the only thing affecting information is aperture? No, I do not believe we can. If we keep f-ratio constant and scale the aperture we are scaling the focal length. This will allow us to gain spatial resolution up to the point at which the data coming through the cover-glass have revealed all they can and there is nothing left to be pulled from it (owing to the inherent blur). Going beyond that and we are not actually gaining information in the image (there is no more detail to observe). We are clearly losing information as well by restricting our FOV. We're also losing information by dropping the photon count in each pixel. By dropping the photon count in each pixel, we're getting closer not only to the read noise but also to the shot noise. This is what is missing in most discussions of f-ratio and SNR.

Consider Figure 2-1 above again. On the left, we see the SNR is clearly much, much better for the longer exposure. Sure, the short one is closer to the read noise, but both the two longer exposures are far away from it (and, in truth, with the 7 e- of read noise in the camera being about 9.4 ADU, even the 0.001s exposure is well above this as the bias-corrected mean value here was 157 ADU). The difference in the SNR comes down to the photon count per pixel here and will follow a sqrt(N) function. That function looks like this here in Figure 2-2:

Figure 2-2



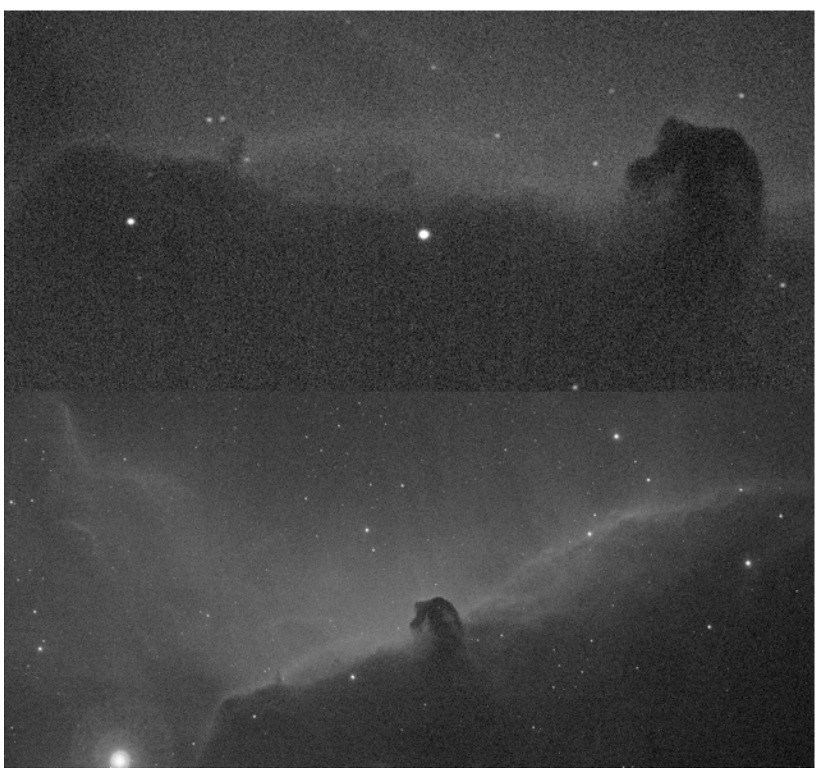
The more photons you put in that well to measure, the higher the SNR, even when read noise is zero (blue line). Read noise tacks on an additional penalty (green line) and skyglow does even more (orange line). Now, once the SNR gets above a certain level, our eye perceives it as a relatively clean image and it's not like doubling the SNR will always make an image look "twice as good". But, people do need to realize that getting a decent number of photons into the well is important for a clean image, regardless of how much spatial information is in there. ***Pixel SNR matters for extended objects***.

One other thing we can get from Figure 2-2 here is why read noise can be so important for things like line filter work. The read noise is the only difference between the green and blue lines and both of these assume there is no skyglow (not too far from the truth for narrowband imaging). If you're trying to reach a certain SNR to make the image look reasonable to the eye, it's going to take a lot more photons to do so with the read noise than without. Here, the read noise is a significant part of the noise. The target photon count and skyglow photon count are both low, so they're not contributing very much and the read noise is a constant penalty. Add some skyglow though, and it quickly swamps out the read noise. In the red line, I've added in a decent bit of skyglow but have removed the read noise. As you can see the difference between it the same thing with read noise is approximately nil.

An Imaging Example
Back in Part 4, I showed some images from Mark Keitel showing an f/7.8 vs. f/5.9 example that, to my eye, showed a real win in SNR for the f/5.9 shot. That was a nice, controlled example. Here, I'll present data with less control, but that some may find striking nonetheless.

A year or so ago, I went out and shot some 5-minute test frames in H-alpha of the Horsehead with an 8" f/5 Antares Newtonian (with a Paracorr) and with my 4" f/4 Borg 101 ED APO. They were shot one right after the other on my QSI 540. Now, the Newt was running at about 1150 mm of focal length (the Paracorr adds 15%), and the refractor was at 400 mm. The Newt gathers 4x as many photons and should have more "information" if aperture is all there is to this. No post-processing was done other than simple stretching. The two images are here in Figure 2-3.

Figure 2-3



Now, I don't know about you, but I'm seeing a lot more detail (or at least as much) in the little 4" f/4 scope than I am in the 8" f/5.75. It's also a lot cleaner. It's one I'd share with someone rather than one I'd go back and re-shoot (e.g., with a longer exposure duration to compensate for the change in the photon flux).

OK, so what is going on here? Well, let's stick a few numbers on this (note, these are updated from my blog entry from back then to fix a few things). The 8" Newt does capture 4x as many photons through it's front end, but it's not like they all hit the CCD. The 92% reflectivity mirrors make it such that only 85% of the light gets into the focuser drawtube. Well, that minus a bit for the central obstruction. Toss in some light loss in the Paracorr and you're down to about 78% of the photons hitting the CCD (quick guess of 98% transmission for each of the elements inside the Paracorr). The Borg won't be perfect either. It's a doublet with a reducer/flattener to get it to f/4 and it'll run at about 92% total throughput. Run the math through here and when we account for the focal length differences as well, we get to the answer that each CCD well is getting hit by only 35% as many photons with the Newt ahead of it relative to with the refractor ahead of it. Put another way, the little 4" f/4 is cramming 2.88x as many photons in each CCD well. Looking back at Figure 2-2, it's no wonder the image looks cleaner.

Can this be true? Well, we can measure the same area in both images. My camera's bias signal is about 209 in this area. I measured the mean intensity in a 10x10 region using Nebulosity's Pixel Info tool for three areas right around and in the Horsehead. On the Borg, they measured 425, 302, and 400. On the Newt, they measured 287, 254, and 278. Now, if we pull out the 209 for the bias signal we have 216, 93, and 191 vs. 78, 45, and 69. If we calculate the ratios, we have 2.76x, 2.07x, and 2.77x. Average these and we're at 2.5x. The back of envelope math said it should be 2.88x. That's pretty darn close for envelope math. But, this is again, just simple geometry. We're covering more sky per pixel on the 400 mm scope than on the 1150 mm scope and our aperture, while bigger, hasn't scaled enough to compensate. Guess what would have scaled enough to compensate? An 11.3" f/4 scope.

Again, it's not that the 4" APO will always win. It'll lose out on maximal spatial detail here as 400 mm is undersampling even my skies. A 100" f/4 scope won't produce the same image as a 1" f/4 scope. The target will be a lot bigger and you'll have more detail. But, the brightness (density, photon count, ADU off your CCD, etc.) for an extended object will be the same. ***Photon count for an extended object is driven by f-ratio. Image scale is driven by focal length. Want more resolution at the same pixel-wise SNR? Boost the aperture but keep the same f-ratio. Want more SNR in your images and you're either willing to trade some spatial information OR you're already asking for more spatial resolution than your conditions will give? Drop the f-ratio.*** My skies won't support running more than about 1500 mm on even the best of nights. Most nights, I won't see real resolution improvements on this versus 1000 mm. In addition, a sharp but noisy image is unlikely to impress and running at lower f-ratios will pack more photons onto that CCD well boosting the accuracy of our estimate of that patch of sky's true value, my measure of SNR.

Summary

To re-iterate - as I raise the focal length up to, call it about 1500 mm, I will be increasing the potential spatial information in my images (with my skies, and my 7.4u pixels - your value will vary). Going beyond that and I won't be gaining much, if any, spatial information. Ideally, I'd run each of these focal lengths with as low an f-ratio as possible. Of course, lowering the f-ratio for a constant focal length means increasing the aperture, leading to the notion that aperture rules and leading to the conclusion that we should reach for the biggest scope we can find. But, there are two other sides to this that should be considered. First, as we go up in focal length, passing that point into over-sampling is leading you to lose SNR on extended objects without gaining any actual spatial information. Second, spatial information alone may not be what we're all after. I will gladly trade some amount of spatial information for a cleaner galaxy or more galaxies in my background. For that, photon density per CCD well rather than total number of photons collected matters most. Here is where the f-ratio clearly steps in as it is what determines the photon density (aperture = total photons; focal length = spread of photons; f-ratio = density of photons).

Many amateurs routinely run in photon-poor conditions. We use small pixels on DSLRs. We use one-shot color cameras with their built-in filters or we add our own filters that cut the photon counts (especially, but not limited to line-filters). We grab f/8 and f/10 scopes that spread the photons thin and we image for a minute or so. All of these conspire to cut the photons per pixel down which cuts the SNR of our extended objects. Run that scope at a lower f-ratio (which, yes, will make each pixel cover more sky as that's the whole point) and you'll image that nebula better.

Myth 3: There is something about film that made f-ratio apply that isn't true of CCDs (or there is something about CCDs that makes the f-ratio not apply).

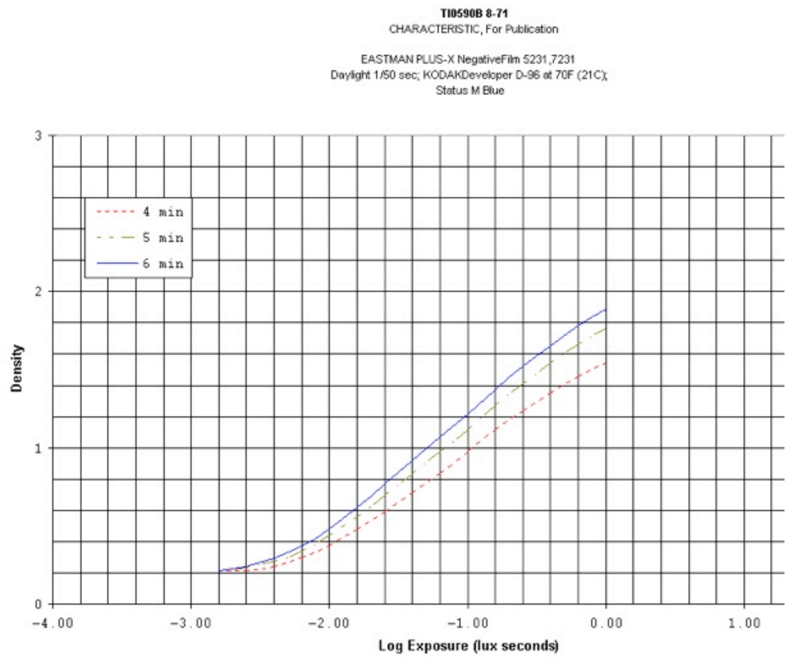
Hopefully, the above is at least making it clear that the f-ratio does matter with CCDs for our DSOs. In some ways, this needn't be covered here, but it is something I think it's worth setting the record straight on. Much has been made about "reciprocity" in film. First, let's get the terms clear.

"Reciprocity" refers to the reciprocal relationship between exposure duration and intensity. That is, that the density on film is equal to the exposure time multiplied by the intensity. Inherent in this is the notion that film is a linear medium while inside a certain range. "Reciprocity failure" is the breakdown of this. With very low flux rates, film became non-linear. It takes several dozen photons to get a halide crystal to form a latent image and if they don't arrive within a certain amount of time, the captured photons are lost. This, of course, hits astrophotographers (who use film) extensively. But, this has nothing to do at all with the notion from general photographers that doubling the exposure duration is equal to one f-stop. These photographers quite correctly used the reciprocity rule as they had enough photon flux to ignore reciprocity failure.

Now, the film response isn't purely linear and it's not always easily characterized, but it certainly can be quite linear over a certain range. Here, in Figure 3-1, we have a plot of Kodak Plus-X film at various exposure durations and at three different development times.

The "knee" on the left there is the reciprocity failure, but beyond -2.0 or so on the graph, we're looking nice and linear. Film can be linear and the use of a reciprocal relationship between exposure duration and f-ratio by photographers isn't the result of any oddness of film. Rather, it's the result of the fact that for much of this graph, things are linear and we can swap out one for the other. That is the definition of reciprocity. Again, where it breaks down is in the very low flux situations (the left side of this graph), typically only encountered by astrophotographers, high-speed photographers, and microphotographers (all with low flux counts). In this reciprocity failure area, you fail to record the target. So, getting the flux rate above this is crucial to recording the target. Lowering the f-ratio will, of course, get you up off this knee better. But, f-ratio here is helping you get out of the non-linear zone. The golden rule of f-ratios changing the density of the recorded image has nothing to do with this zone and has everything to do with the nice linear zone. One f-stop is equivalent to doubling the flux, a relationship that only holds when things are linear.

Figure 3-1



Now, the big difference between CCD and film here (apart from an overall difference in sensitivity) is the fact that the CCD response is not only linear in the meat of the range, but it is also linear on the low end. Here is a plot from Kodak's spec sheet for one of their sensors. Numerous similar examples exist on the web (e.g., a Kodak KAF-1602E's response from an SBIG ST8E measured by the Astrophysics and Space Science Institute at UMBC, and a sample spec sheet from Apogee):

Figure 3-2

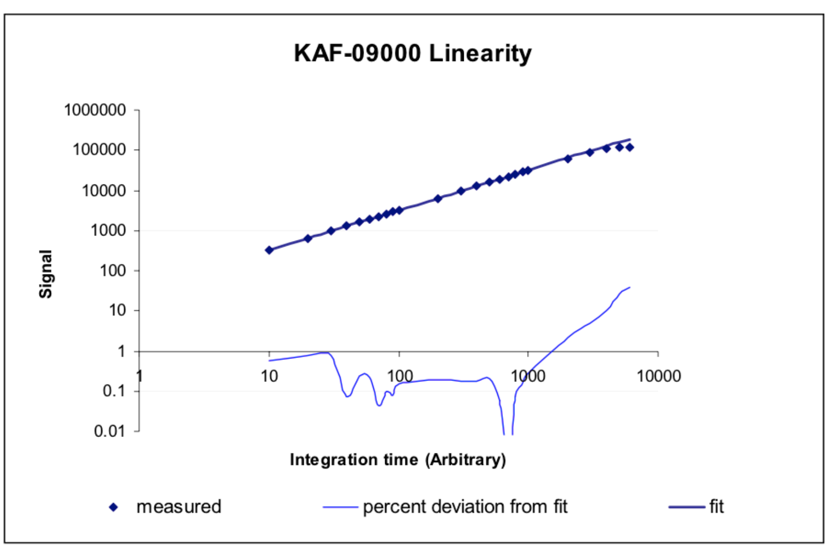


Figure 9: Linearity

You’ll notice that the bottom end is nice and linear. The top end starts to show saturation (much as film does). But again, it is a misnomer to say that “CCD’s are linear and film isn’t.” Film can be quite linear and typically is very linear within a certain range (the range photographers typically use and the range the f-ratio rule is made for). Both show “reciprocity” and that’s a good thing. Above and below that range it goes non-linear. For many of our CCDs, above a certain range, they also go non-linear. The great thing is that not only do they not suffer from reciprocity failure, but they are far more sensitive overall than film. However, in the linear range, doubling the photon flux (e.g., by dropping by one f-stop) doubles the density on the film and doubles the number of photons captured by the CCD.

With this, we have likely reached the conclusion of this series on SNR. Others may wish to continue with topics like the role that f-ratio plays on star shape, the role for imperfections in CCDs, a more thorough treatment of the role of f-ratio on stellar sources, or a more thorough treatment of object SNR (or, perhaps more appropriately, a treatment of "image quality" - how well all of the information from that portion of the sky is being recorded.) I hope from my modest efforts on the topic, that some have learned a few things along the way. Whether you enjoyed the ride, found it obtuse and confusing or thought it myopic, I wish you clear, steady, and dark skies so that you can enjoy fishing for those photons.

Craig