ORF 245 Fundamentals of Statistics
Chapter 8
Parameter Estimation

Robert Vanderbei

Fall 2014

Slides last edited on November 18, 2014

Method of Moments (MoM)

Let X be a random variable. The *moments* of X are defined as

$$\mu_1 = \mathbb{E}(X), \qquad \mu_2 = \mathbb{E}(X^2), \qquad \mu_3 = \mathbb{E}(X^3), \qquad \ldots$$

If $X_1, X_2, \ldots, X_n$ are iid random variables having the same distribution as X, then the *sample moments* are defined as

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \qquad \hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{n} X_i^3, \qquad \ldots$$

The sample moments can be thought of as estimates of the true moments.

Suppose we know that X has a distribution of a certain type (such as Poisson, or Normal, etc.) but that the values of the particular parameters that characterize that distribution are unknown.

If the distribution involves just one unknown parameter (for example, the Poisson distribution), then we can write a single equation in a single unknown that we can use to solve for an estimator of the unknown parameter.

If the distribution involves two unknown parameters (for example, the Normal distribution), then we can employ the same idea using the first two moments.

And so on.

MoM for Poisson

The Poisson distribution involves a single parameter λ.

As we've shown before,

$$\mathbb{E}(X) = \lambda$$

Hence, the MoM suggests using

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

as an *estimator* for λ.

Normal MoM

The Normal distribution involves a two parameters μ and σ^2.
As we've shown before,

$$\mu_1 = \mathbb{E}(X) \quad = \mu$$

and

$$\mu_2 = \mathbb{E}(X^2) = \sigma^2 + \mu^2$$

Thinking of μ_1 and μ_2 as known (or, at least, "estimatable"), we can think of this as two equations in two unknowns (μ and σ^2) which can be solved

$$\begin{aligned} \mu &= \mu_1 \\ \sigma^2 &= \mu_2 - \mu_1^2 \end{aligned}$$

From these formulas, we get the following MoM estimators of μ and σ^2:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \end{aligned}$$

Maximum Likelihood: An Alternative to MoM

Consider a random variable X with a probability density (or mass) function $f(x)$.

Suppose that this density function depends on a parameter. Let's call it θ and write the density function as $f(x|\theta)$.

If we think of x as a parameter and θ as the variable, then this function is called the *likelihood* that θ is the correct parameter value given an observation x:

$$\text{lik}(\theta) = f(x|\theta)$$

If several iid observations of random variable X are taken, then the joint density function is the product and so the likelihood function is also a product:

$$\text{lik}(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

The *maximum likelihood estimator* is that choice of θ that makes the likelihood the largest.

Maximizing the likelihood involves setting a derivative to zero. Since, derivatives of products are tedious to compute, a standard trick is to maximize the logarithm of the likelihood

$$l(\theta) = \log(\text{lik}(\theta)) = \sum_{i=1}^{n} \log(f(x_i|\theta))$$

MLE for Poisson

The density function for a Poisson random variable is

$$f(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, 2, \ldots$$

Hence, the log-likelihood function associated with n independent observations is

$$l(\lambda) = \sum_{i=1}^{n} \left(x_i \log(\lambda) - \log(x_i!) - \lambda \right)$$

We maximize by taking the derivative and setting it to zero:

$$\frac{d}{d\lambda} l(\lambda) \; = \; \sum_{i=1}^{n} \left(\frac{x_i}{\lambda} - 1 \right) \; = \; 0$$

Solving for λ, denoting the solution by $\hat{\lambda}$ and using *a priori* (i.e., probabilistic) upper-case random-variable notation we see that

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

In words, the MLE estimator for parameter λ is just the sample mean (same as MoM).

MLE for Normal

The density function for a Normal random variable is

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty$$

The log-likelihood function associated with n independent observations is

$$l(\mu,\sigma) = \sum_{i=1}^{n} \left(-\log\left(\sqrt{2\pi}\sigma\right) - \frac{(x_i-\mu)^2}{2\sigma^2} \right)$$

Differentiating, we get

$$\begin{aligned} \frac{\partial l}{\partial \mu}(\mu,\sigma) &= \sum_{i=1}^{n} \frac{(x_i-\mu)}{\sigma^2} \\ \frac{\partial l}{\partial \sigma}(\mu,\sigma) &= \sum_{i=1}^{n} \left(-\frac{1}{\sigma} + \frac{(x_i-\mu)^2}{\sigma^3} \right) \end{aligned}$$

MLE for Normal – Continued

Setting the first partial derivative to zero and solving for $\hat{\mu}$, we get

$$\sum_{i=1}^{n} \frac{(x_i - \hat{\mu})}{\sigma^2} = 0 \qquad \Longrightarrow \qquad \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

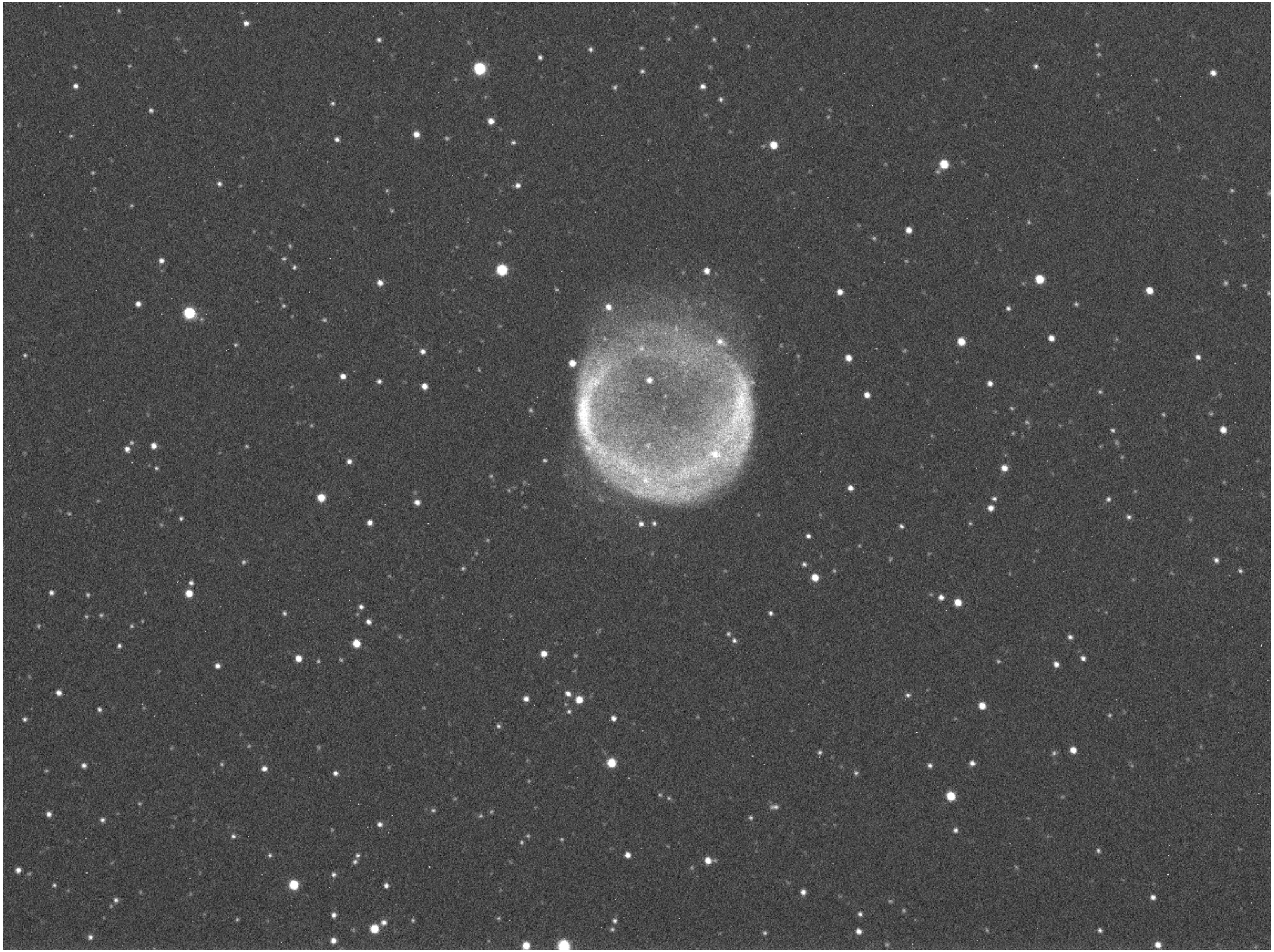
Next, setting the second partial derivative to zero and solving for $\hat{\sigma}$, we get

$$\begin{aligned}
\sum_{i=1}^{n} \left(-\frac{1}{\hat{\sigma}} + \frac{(x_i - \hat{\mu})^2}{\hat{\sigma}^3} \right) = 0 \qquad & \Longrightarrow \qquad \sum_{i=1}^{n} \left(-\hat{\sigma}^2 + (x_i - \hat{\mu})^2 \right) = 0 \\
& \Longrightarrow \qquad -n\hat{\sigma}^2 + \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = 0 \\
& \Longrightarrow \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2
\end{aligned}$$

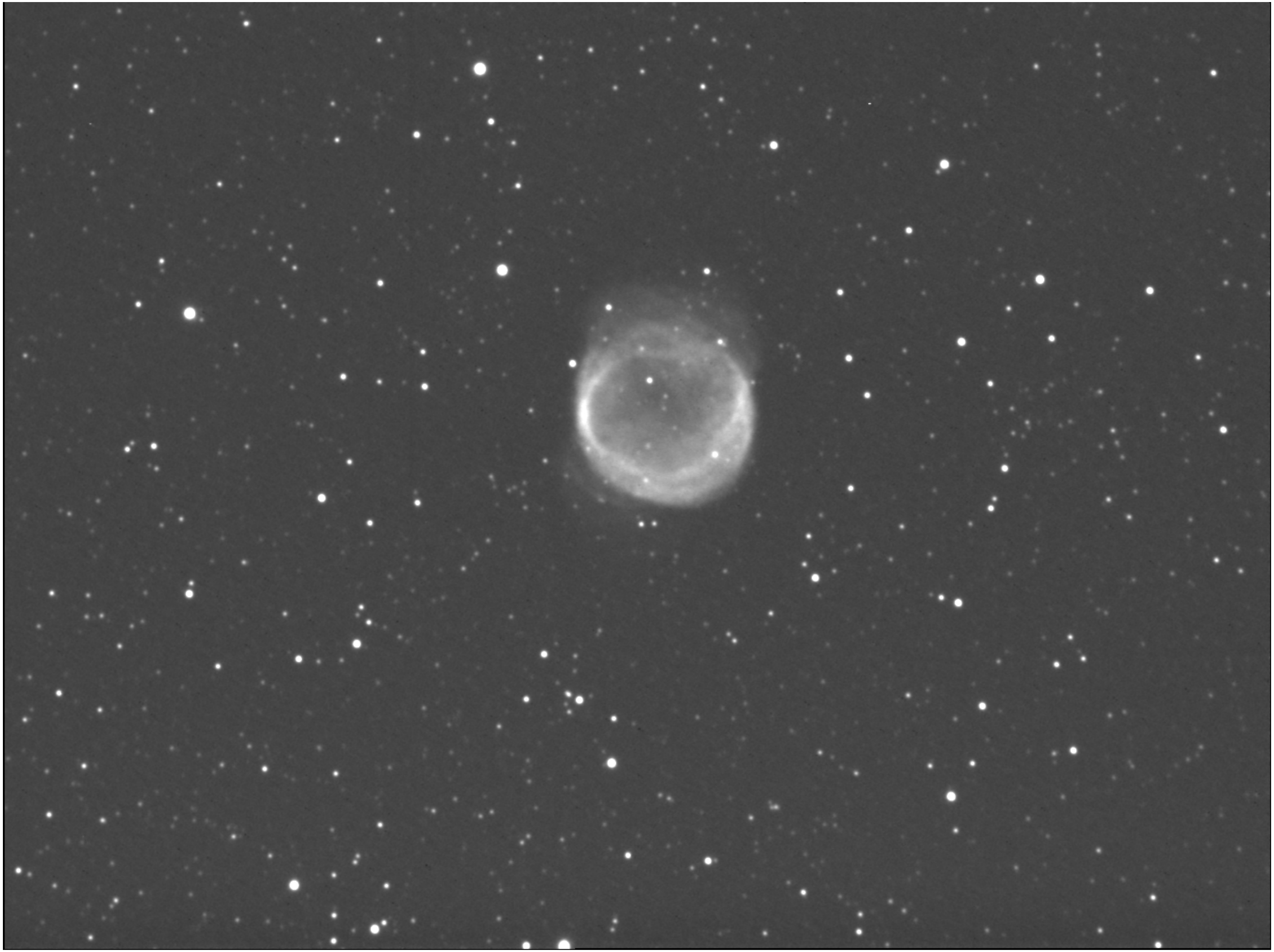
Conclusion: Same answer as was obtained by MoM.

Example: Photography = Counting Photons

NGC 6781 – Red Channel – One Frame



NGC 6781 – Red Channel – 37 Frames



Method of Moments: Digital Camera

For a given exposure time, the number of photons that land on a particular pixel of a digital camera can be perfectly modeled as a Poisson random variable X. The true intensity of the light at this pixel is proportional to the (unknown) mean λ of this random X. The digital camera provides a number to the image file that has the form

$$Y = \alpha + \gamma X$$

The unknown constants α and γ are called the *bias* and the *gain*, respectively.

Using the fact that X is Poisson with parameter λ, it is easy to compute that

$$\mathbb{E}(X) = \lambda, \qquad \mathbb{E}(X(X-1)) = \lambda^2, \qquad \text{and} \qquad \mathbb{E}(X(X-1)(X-2)) = \lambda^3$$

From these, it is easy to find the first three moments:

$$\mathbb{E}(X) = \lambda, \qquad \mathbb{E}(X^2) = \lambda^2 + \lambda, \qquad \text{and} \qquad \mathbb{E}(X^3) = \lambda^3 + 3\lambda^2 + \lambda$$

And, from these it is easy to derive formulas for the moments of Y:

$$\begin{aligned}
\mathbb{E}(Y) &= \alpha + \gamma\lambda \\
\mathbb{E}(Y^2) &= \mathbb{E}(\alpha+\gamma X)^2 = \mathbb{E}(\alpha^2 + 2\alpha\gamma X + \gamma^2 X^2) = \alpha^2 + 2\alpha\gamma\lambda + \gamma^2(\lambda^2+\lambda) \\
\mathbb{E}(Y^3) &= \mathbb{E}(\alpha+\gamma X)^3 = \quad \cdots \quad = \alpha^3 + 3\alpha^2\gamma\lambda + 3\alpha\gamma^2(\lambda^2+\lambda) + \gamma^3(\lambda^3+3\lambda^2+\lambda)
\end{aligned}$$

If we assume that we take n independent digital pictures, then we will have n realizations, $y_1, y_2, \ldots, y_n$, of independent random variables with the distribution of Y.

We will use the sample moments as estimates of the true moments:

$$\mathbb{E}(Y) \approx \frac{1}{n}\sum_i y_i, \qquad \mathbb{E}(Y^2) \approx \frac{1}{n}\sum_i y_i^2, \qquad \mathbb{E}(Y^3) \approx \frac{1}{n}\sum_i y_i^3$$

With these three approximations, the three equations on the previous slide can be viewed as *three equations* in *three unknowns*: α, γ, and λ. Unfortunately, these equations are nonlinear. But, with a little tinkering, they can be solved:

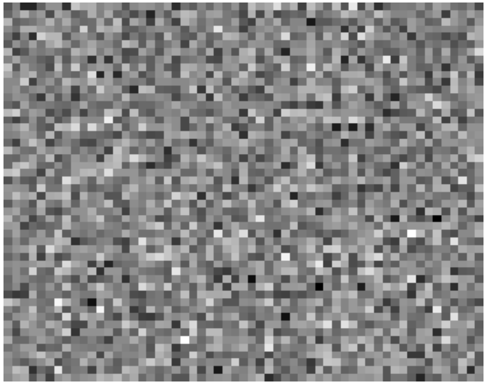
$$\begin{aligned} \gamma &= \frac{\zeta^3}{\sigma^2} - 3\mu \\ \lambda &= \frac{\sigma^2}{\gamma^2} \\ \alpha &= \mu - \gamma\lambda \end{aligned}$$

where

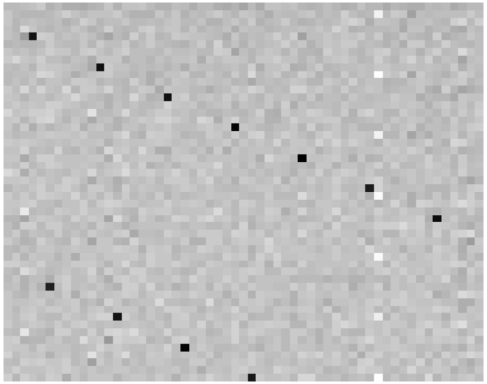
$$\mu = \mathbb{E}(Y), \qquad \sigma^2 = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2, \qquad \text{and} \qquad \zeta^3 = \mathbb{E}(Y^3) - (\mathbb{E}(Y))^3$$

A Very Short Exposure of a “Flat” Field

A cropped single image. Highly enlarged to make each pixel easy to see. Highly random from pixel to pixel.



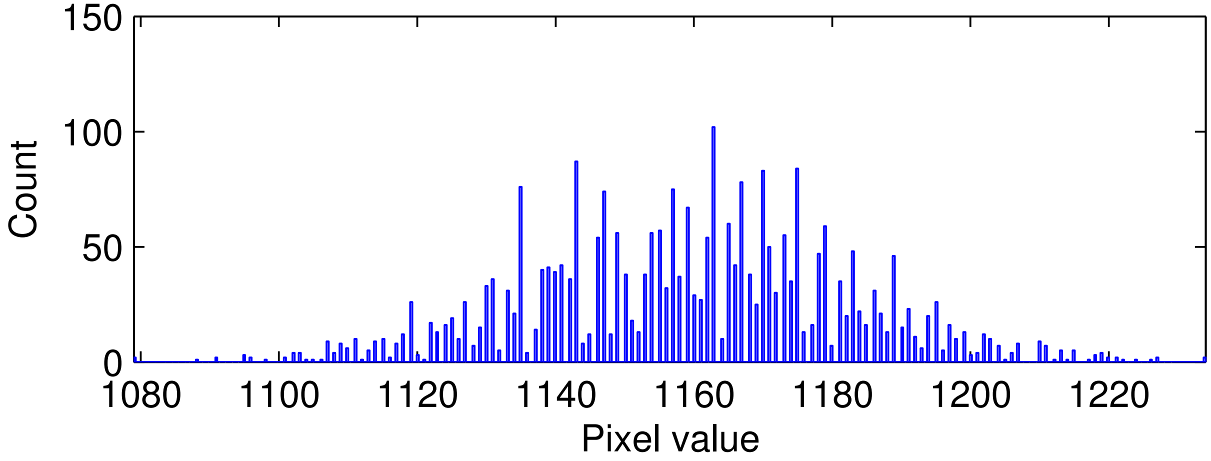
Average of 666 frames. Much smoother. Some “dead” pixels are now evident (and some “warm” ones too).



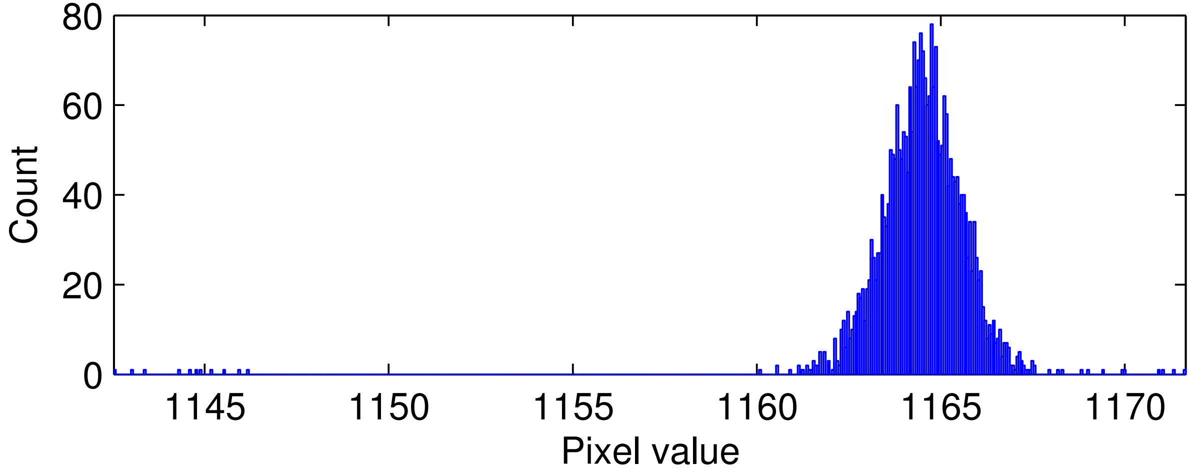
Poisson: The Movie

Associated Histograms

One frame.



Average of 666 frames. The 11 dark pixels are *outliers*.



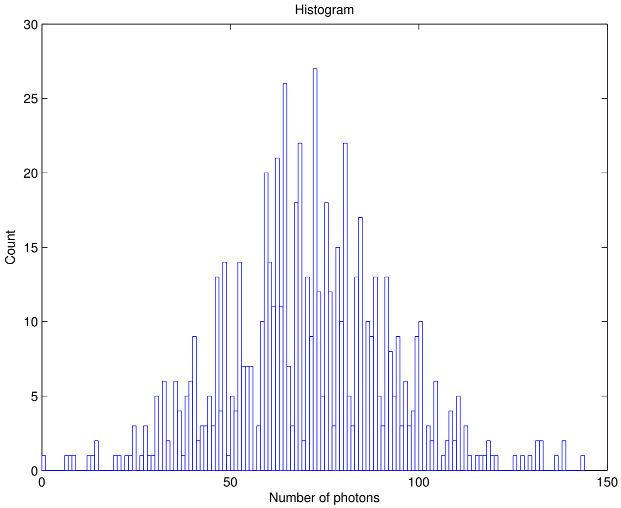
Focus on Just Three Specific Pixels

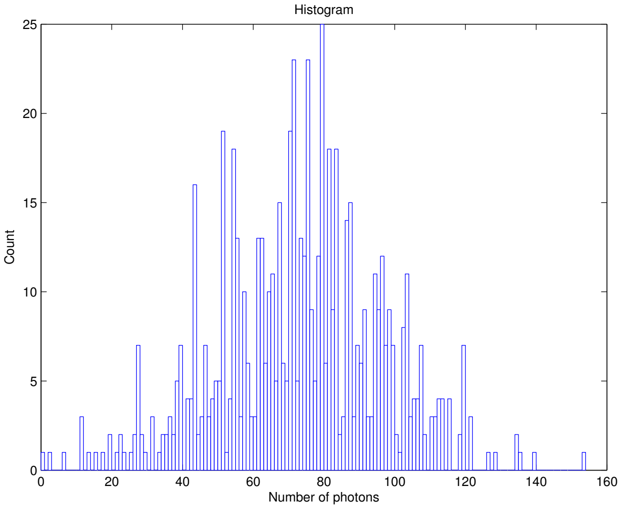
$$(25, 25)$$
$$(25, 30)$$
$$(30, 30)$$

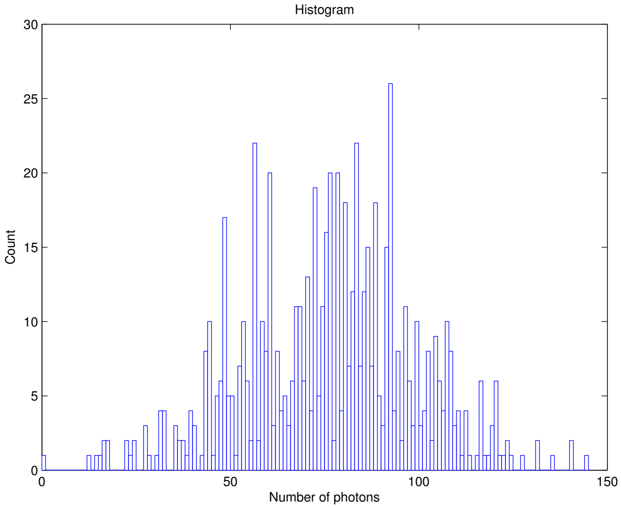
Matlab command for making a histogram

having 100 bins using data stored in `x`:

```
histogram(x,100);
```



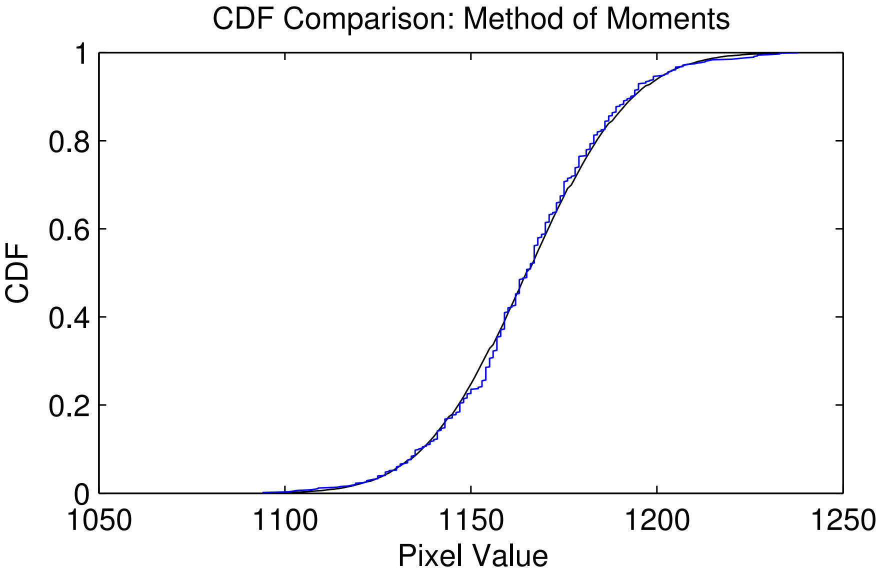




Results

For pixel $(25, 25)$:

$$\alpha = 223, \qquad \gamma = 0.525, \qquad \lambda = 1797$$



Averaging over all pixels in the image, we get

$$\alpha = 210 \pm 20, \qquad \gamma = 0.560 \pm 0.049, \qquad \lambda = 1708 \pm 116$$

Final Comments on MoM for Digital Camera

If the bias, α, is assumed known (from separate independent analysis), then we have just two unknown parameters, γ and λ, and so we need only the first two moments. The algebra is much easier.

The answer is:

$$\gamma = \frac{\sigma^2}{\mu - \alpha}$$

$$\lambda = \frac{(\mu - \alpha)^2}{\sigma^2}$$

This is how photographers actually do it.

Maximum Likelihood: Digital Camera

Recall that $Y = \alpha + \gamma X$, where X is a Poisson random variable with parameter λ.

It is easy to compute the density function for Y:

$$f_Y(y) = P(Y = y) = P(X = (y - \alpha)/\gamma) = \frac{\lambda^{(y-\alpha)/\gamma}}{((y-\alpha)/\gamma)!} e^{-\lambda}$$

(of course, we assume that $(y - \alpha)/\gamma$ is an integer.

If we have several independent observations, then the joint density function is just the product:

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{\lambda^{(y_i-\alpha)/\gamma}}{((y_i-\alpha)/\gamma)!} e^{-\lambda}$$

Viewed as a function of α, γ, and λ, this function is called the *likelihood function* and its logarithm is called the *log-likelihood function*:

$$l(\alpha, \gamma, \lambda) = \sum_{i=1}^{n} \left(\frac{y_i - \alpha}{\gamma} \log(\lambda) - \log\left(\frac{y_i - \alpha}{\gamma}! \right) - \lambda \right)$$

Ugh!!!

Example 8.5.A — Hardy-Weinberg Model

In genetics, a particular gene can be either dominant (A) or recessive (a). A person inherits one copy of the gene from each parent and therefore has a genetic makeup consisting of two copies of the gene. Hence, there are three possible *genotypes*:

AA: purely dominate Aa: heterozygotic aa: purely recessive

Suppose that, in the entire population, the dominate gene A occurs with probability p and the recessive gene a occurs with probability $q = 1 - p$. Then, according to the *Hardy-Weinberg* model, the probability that a randomly selected person has a specific genotype is given by

AA	Aa	aa
p^2	$2pq$	q^2

We don't know p. But, we can do genetic testing and determine out of n people that X_1 are purely dominate, X_2 are heterozygotic, and X_3 are purely recessive in this geneotype.

The joint pmf for these random variables is

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{n!}{x_1!x_2!x_3!}(p^2)^{x_1}(2pq)^{x_2}(q^2)^{x_3}$$

A fundamental question is: what function of X_1, X_2, and X_3 is a good estimator of p?

Hardy-Weinberg Model — MOM

In this model, there are three random variables. Hence, there are three first moments:

$$\mathbb{E}(X_1) = np^2, \qquad \mathbb{E}(X_2) = 2npq, \qquad \mathbb{E}(X_3) = nq^2$$

But, there is only one parameter: p.

Since each of these random variables can be thought of as a Binomial random variable and since a Binomial random variable is a sum of Bernoulli random variables, the *method of moments* can be used to find estimators for p.

If we use the equation for the mean of X_1, we get the estimator

$$P_1 = \sqrt{X_1/n}.$$

On the other hand, if we use the equation for the mean of X_2, we get two estimators

$$P_2 = \frac{1 \pm \sqrt{1 - 2X_2/n}}{2}.$$

Lastly, if we use the equation for the mean of X_3, we get the estimator

$$P_3 = 1 - \sqrt{X_3/n}.$$

Each of these estimators uses only one of the three variables. Can we do better? Yes...

Hardy-Weinberg Model — MLE

The log-likelihood function is given by

$$l(p) = c + x_1 \log(p^2) + x_2 \log(2p(1-p)) + x_3 \log((1-p)^2).$$

Setting the derivative to zero, we get

$$\frac{\partial l}{\partial p}(p) = x_1 \frac{2}{p} + x_2 \left(\frac{1}{p} - \frac{1}{1-p} \right) - x_3 \frac{2}{1-p} = 0.$$

Multiplying both sides by $p(1-p)$, we get

$$2x_1(1-p) + x_2(1-2p) - 2x_3 p = 0.$$

Solving for p, we get

$$p = \frac{2x_1 + x_2}{2x_1 + 2x_2 + 2x_3}$$

And, since $x_1 + x_2 + x_3 = n$, we finally get that the *maximum likelihood estimator* of p is

$$\hat{P} = \frac{2X_1 + X_2}{2n} = \frac{1}{2} + \frac{X_1 - X_3}{2n}$$

Note that the first formula is highly intuitive!

Hardy-Weinberg Model — Real Data

Here's some real data (see textbook for source):

AA	Aa	aa
342	500	187

Using this data, we can make the following MOM estimates for p (and q):

$$\begin{aligned} \hat{p}_1 &= \sqrt{345/1029} = 0.5790 \\ \hat{p}_2 &= \frac{1 \pm \sqrt{1 - 2 \times 500/1029}}{2} = \begin{cases} 0.5839 \\ 0.4161 \end{cases} \\ \hat{p}_3 &= 1 - \sqrt{187/1029} = 0.5737 \end{aligned}$$

Using the data, we can make the following MLE estimate for p (and q):

$$\hat{p} = \frac{1}{2} + \frac{342 - 187}{2 \times 1029} = 0.5753, \qquad \hat{q} = 1 - \hat{p} = 0.4247$$

A natural question arises: How good are these estimates?

See next slide.

Bootstrap

If we had several sets of independent data, we could compute several values of $\hat{p}$ and compute the average and the standard deviation of these values.

Unfortunately, we only have the one data set shown on the previous slide.

But, we do know the theoretical distribution underlying this data and that's extremely helpful.

We can assume that $\hat{p} = 0.5753$ is reasonably close to the correct answer and we can then artificially generate thousands of new data sets using this value of p.

The new data sets will be random—the values of x_1, x_2 and x_3 will vary.

We can then compute many different estimates of $\hat{p}$ based on these simuations.

The average value will be very close to 0.5753.

But, the key is that we can compute a standard deviation which will give us an idea how precise our estimate $\hat{p} = 0.5753$ is.

Bootstrap — Continued

Here's Matlab code to run bootstrap $10{,}000$ times:

```
    p = 0.5753;
    m = 10000;
    n = 1029;
    M = random('unif',0,1,[n m])<p;         % male parent
    F = random('unif',0,1,[n m])<p;         % female parent
    G = M + F;                             % 2=dominate, 1=heterozygote, 0=recessive
    x1 = sum(G == 2);
    x2 = sum(G == 1);
    x3 = sum(G == 0);
    p = 0.5 + (x1-x3)/(2*n);
    mean(p)
    std(p)
```

The answers for $\hat{p}$ are:

$$\text{mean(p)} = 0.5753, \qquad \text{std(p)} = 0.0110$$

Hence, an approximate 95% confidence interval for the true value of p is

$$0.5753 - 2 \times 0.0110 \ \leq \ p \ \leq \ 0.5753 + 2 \times 0.0110$$

Values for the mean and standard deviation of $\hat{p}_1$ and $\hat{p}_3$ are also easy to compute:

$$\begin{aligned} \text{mean(p1)} &= 0.5751, \qquad \text{std(p1)} = 0.0129 \\ \text{mean(p3)} &= 0.5754, \qquad \text{std(p3)} = 0.0142 \end{aligned}$$