*ASTR 2401*

# A Quick Guide to Statistics

Observational Astronomy

# Labs This Week

- Observing Labs Start at 7:00 - **PROJECTS**
  - **Bring your observing plan and finder charts**
  - 8:00 for alternate lab on campus
  - I think we will go out tonight and hope, Thursday looks grim.

# New Lab Six

- On your own computer, download AstroImageJ from http://www.astro.louisville.edu/software/astroimagej/installation_packages/

- Install AstroImageJ

- Download the User Guide, the tutorial in Chapter 10 will replace Lab Six

- You can download the Tutorial data files a head of time, but they are big (4 GB)! I'll have thumb drives with the files for you in the lab on Thursday.

- Bring your computer to Lab on Thursday

# Basic Quantities

- **Mean (average):**

$$\mu = \frac{1}{N}\sum_i x_i$$

- **Median**: Half of the values are larger, and half of the values are smaller.
  - Robust to outliers
  - If have an even number of points, take the average of the middle two

- **Mode**: The most probable value (one that occurs most often).

*Example: An observation is made in which a single star is observed 7 times. the number of counts in the detect for that star is:*

80, 120, 103, 90, 94, 103,17

Mean: (80+120+103+90+94+103+17)/7 = 86.7

Median: 94

Mode: 103

# Basic Quantities

- Mean (average):

$$\mu = \frac{1}{N}\sum_i x_i$$

- Weighted mean: factors in uncertainty in individual measurements.

$$\mu = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}$$

# Standard Deviation

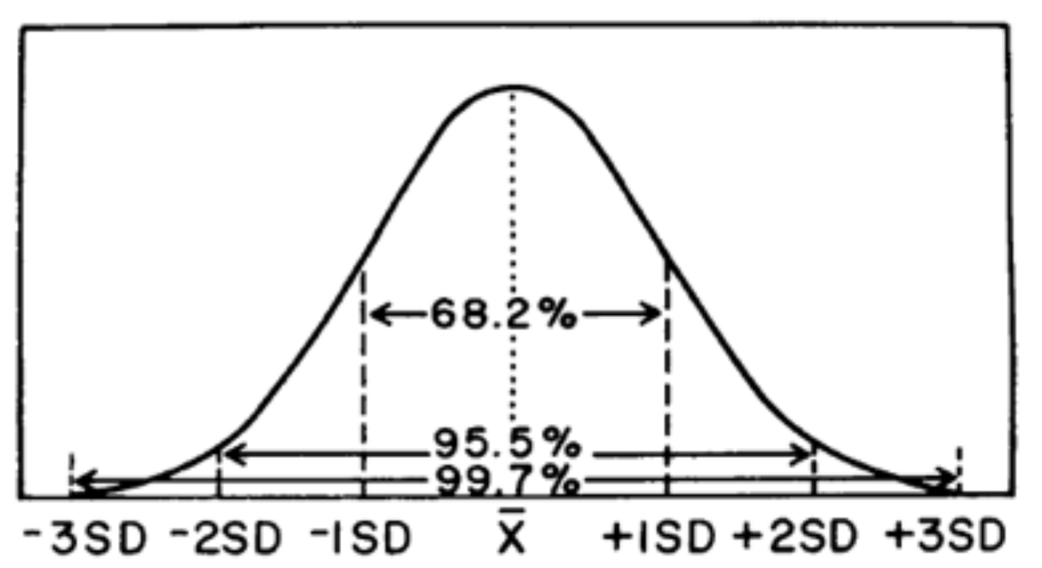
- **Variance**:

$$\sigma^2 \equiv \langle (x-\mu)^2 \rangle,$$

- **Standard Deviation:**

$$\sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$$

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_i (x_i - \bar{x})^2}$$



*Standard deviation is important for determining uncertainties. For a Gaussian distribution,*

$\Rightarrow$ *68.3% probability that the true values is within 1σ of the mean*

$\Rightarrow$ *95.4% probability that the true values is within 2σ of the mean*

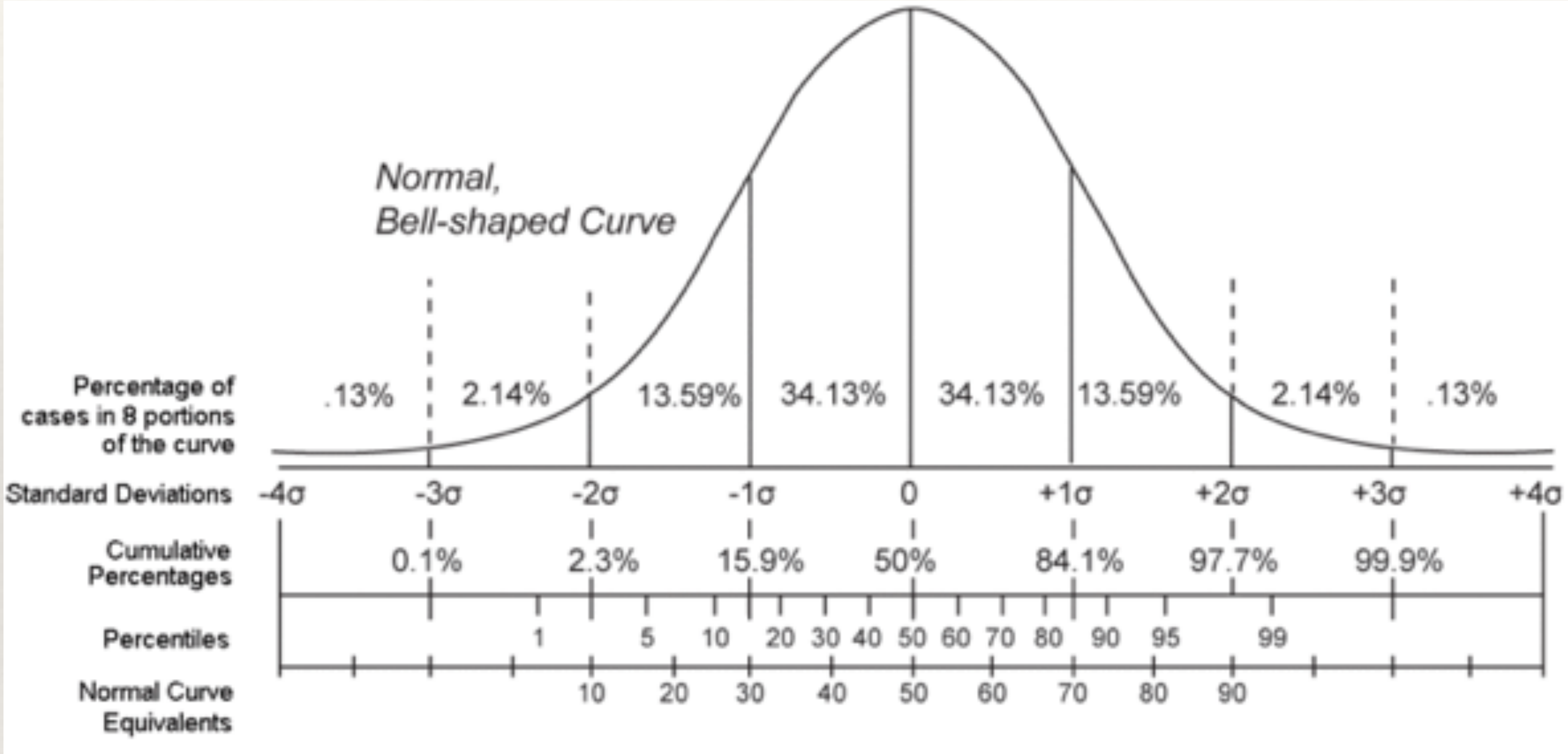
$\Rightarrow$ *99.7% probability that the true values is within 3σ of the mean*

$$P = erf(N/\sqrt{2})$$

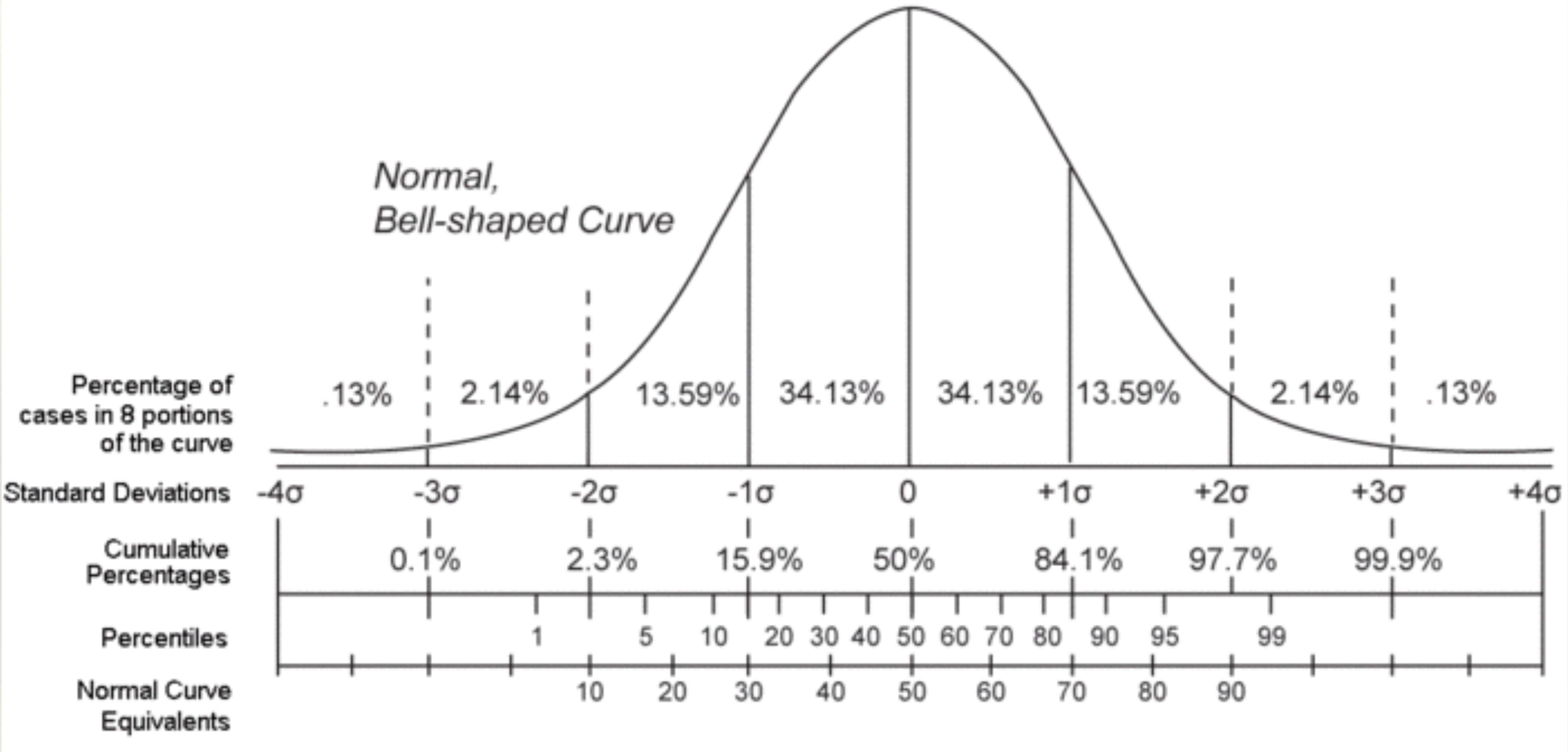
*More generally, the probability that a point lies within N σ of the mean is*

# Gaussian (Normal) Distribution

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



# Gaussian (Normal) Distribution



Standard deviation is important for determining uncertainties. For a Gaussian distribution,

⇒ 68.3% probability that the true values is within 1σ of the mean

⇒ 95.4% probability that the true values is within 2σ of the mean

⇒ 99.7% probability that the true values is within 3σ of the mean

# Poisson Distribution

The Poisson distribution is of fundamental importance in astronomy. Put simply, the Poisson distribution asks how many times an event is likely to happen in a given amount of time. Examples of applications in the real would include:

How many bagels will be sold at Einstein's Bagels today?
How many cell phones will be sold in the US this week?

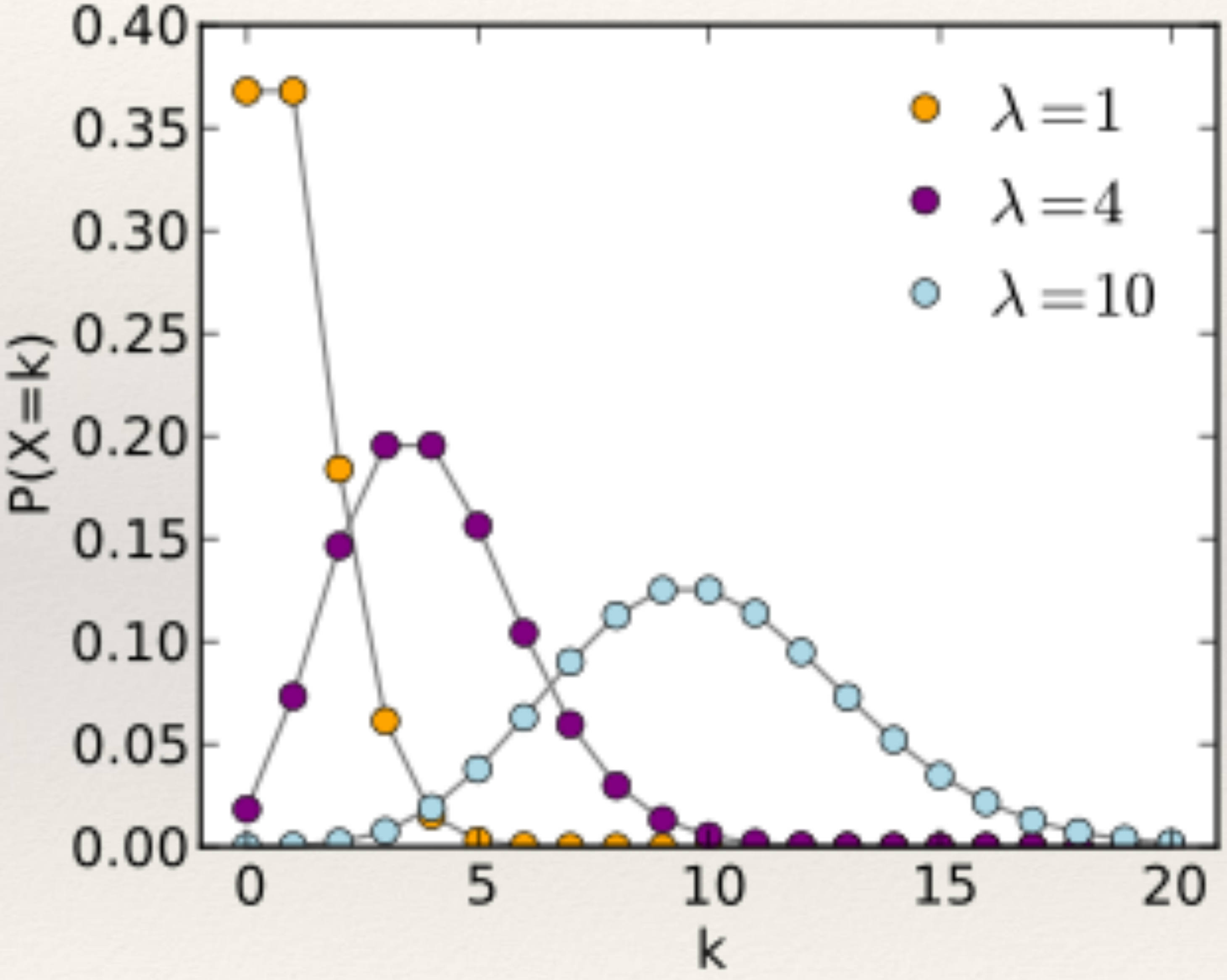
In astronomy, the question is: How many photons will hit my detector while I'm observing an object?

In all cases the answer can be any ***non-negative integer***. The probability of getting a certain number of occurrences is:

$$P(N=k)=\frac{e^{-\lambda}\lambda^{k}}{k!}$$

where $\lambda$ is the expected number of events when averaged over a very long period.

# Poisson Distribution



# Poisson Distribution

Example: Assume that the average number of courses taken per semester by students at TTU is 3 (and for the moment assume that you're allowed to take anywhere from 0 to infinitely many). If you ask someone how many they are taking, if the distribution is Poisson then:

P(0) = 0.02
P(1) = 0.07
P(2) = 0.15
P(3) = 0.20
P(4) = 0.20
P(5) = 0.16
P(6) = 0.10
P(7) = 0.06

$$P(N = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

In the limit of λ becoming very large, the mean and standard deviation for a Poisson distribution approach:

$$\sigma = \sqrt{\lambda} \approx \sqrt{N}$$

For optical astronomy, where you have a lot of photons, these are very good approximations.

# Poisson Distribution Lab
# Data Collection due by Thursday Nov 10th

You'll go to a stoplight, and count the cars that pass through a green light.

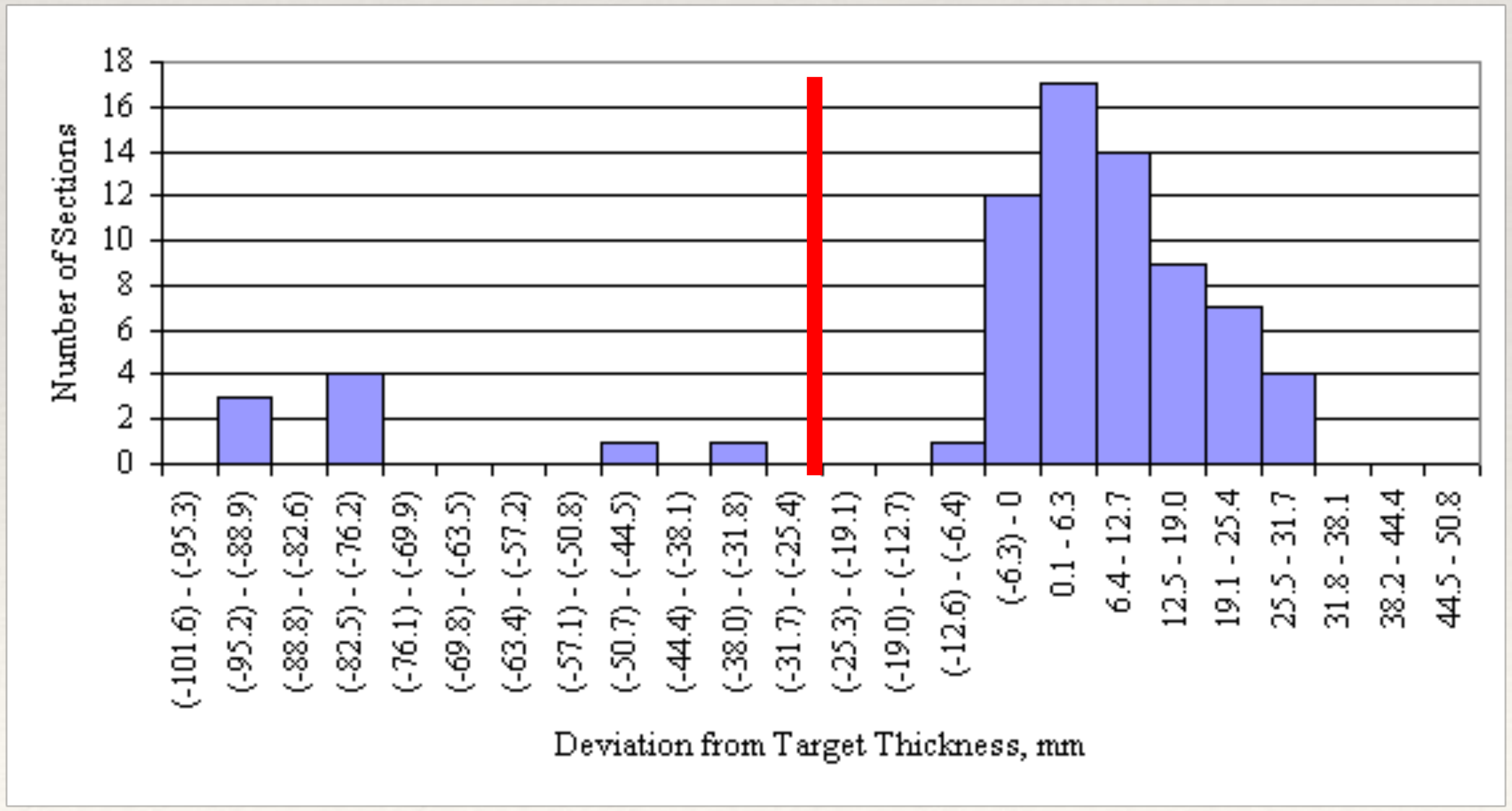
You'll then estimate the rate using Poison Statistics.

This is analogous to how CCDs collect data.

Instructions posted on course website.

# Outlier Rejection

- Iterative ***"Sigma Clipping"*** is a common technique in astronomy for dealing with outliers, particularly when combining images to eliminate artifacts such as cosmic rays. Another example would be in computing the velocity dispersion of a galaxy cluster, for which the "outliers" are galaxies not associated with the cluster.
- Basic method:
  1. Compute the mean (median) and standard deviation.
  2. Reject all points >N σ away from the mean (median) as outliers. A typical value for N might be 5 or 10.
  3. Recompute the mean (median) and standard deviation, and again reject outliers.
  4. Repeat until you are no longer rejecting any points.



From USDOT report on LTPP Pavement Layer Thickness

# Least Squares Fitting

Assume that you have a set of data, and wish to compare it with a model.

Two basic questions:

(1) What are the best fit model parameters?

(2) Is the model a good fit to the data?

A standard approach to the first part is “least-squares” fitting. The basic idea is that for a function f(x), the parameters that minimize the square of the difference between the model and each data point will be the best model parameters. (Squaring leads to positive and negative deviations being treated equally.)

The equation is:

$$S = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

With the best solution coming from the parameters that minimize S. If you are fitting a line, f(x)=ax+b, then this is called linear least-squares fitting. The method though is general.

# Chi-Squared Fitting

Chi-squared minimization, or chi-squared fitting, is a refinement of the least squares method that takes into account the error bar associated with each data point. This additional information lets you answer both questions:

(1) What are the best fit model parameters?

(2) Is the model a good fit to the data?

If each data point is now represented by 3 numbers $(x_i, y_i, \sigma_i)$, where $\sigma_i$ is the error bar,

then the quantity to be minimized is:

$$\chi^2 = \sum_{i=1}^{N} \left( \frac{y_i - y}{\sigma_i} \right)^2$$

Note that now what you are fundamentally doing is taking the ratio of the scatter between the data and model to the error bar and summing this for all the data points. If the model is a good fit to the data, then these should on average be equal to ~1.

# Chi-Squared Fitting

Quantitatively, the way that you answer the question of whether the model is a good fit to the data is by computing a quantity called the “**reduced**” chi-squared,

$$\chi^2_\nu = \frac{\chi^2}{N - M - 1} = \frac{\sum_{i=1}^{N}(y_i - f(x_i))^2}{N - M - 1}$$

where N is the number of data points and M is the number of tunable parameters in your function.

Example: f(x) = a*x+b , and you are fitting 10 data points which each have $(x_i, y_i, \sigma_i)$

In this case N=10, M=2.

If $\chi^2_\nu \sim 1$ → The model is a good fit to the data.

If $\chi^2_\nu >> 1$ → The model is a bad fit to the data, or your errors are underestimated.

If $\chi^2_\nu << 1$ → Your errors are overestimated.