Lecture 2 - Review of Basic Statistics

Mean, Variance, and Standard deviation

For a one-dimensional variable x_i where
$x_i = [x_1, x_2 \dots x_N]$

"i" can be an index in time or space.

The mean of the data is:

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

The mean is the first moment about zero.

Other measures (which are often confused with the mean):

Median: Value in the center of a population (or the midpoint)

Mode: Most frequently occurring value.

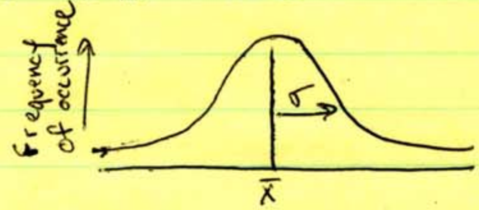
Variance: Second moment about the mean:

$x_i' = x_i - \overline{x}$ → Deviation from the mean.

$$\overline{x'^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i')^2$$

Standard deviation: Square root of the variance (denoted often with the symbol σ)

$$\sigma = \sqrt{\overline{x'^2}}$$



Physically, the standard deviation gives a measure of the spread about the mean. Bigger σ → bigger spread.

~~If data are not normally distributed, then~~
There are also higher moments about the mean:

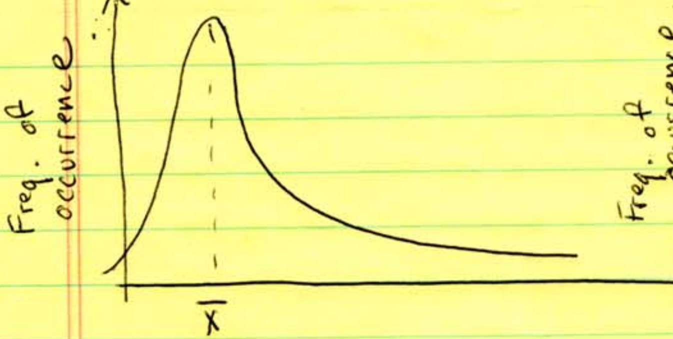
$$m_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^r$$

In this generalization

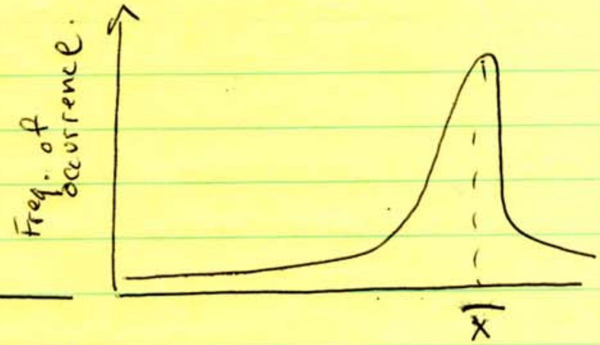
$r = 2$ → Variance
$r = 3$ → Skewness
$r = 4$ → Kurtosis.

Skewness: Indicates degree of asymmetry of the distribution about the mean.

$a_3 > 0$ → longer tail on positive side of mean.
$a_3 < 0$ → negative side.



Positively skewed

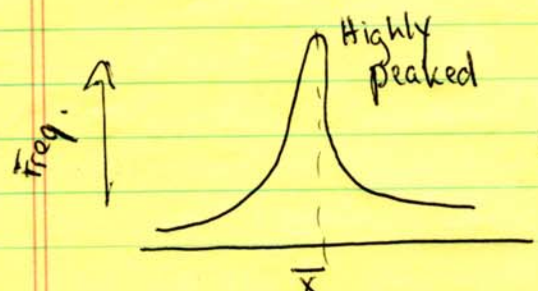


Negatively skewed.

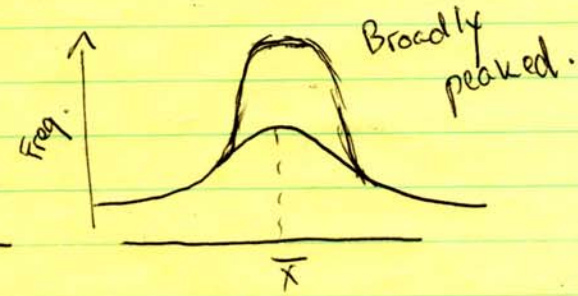
For many types of atmospheric data, a normal distribution (skewness = 0) is a pretty good assumption.

A major exception to this is precipitation, which typically has a positively skewed frequency distribution, especially in arid climates like Arizona. Relatively low mean precipitation but a few intense rain events like monsoon thunderstorms.

Kurtosis: Indicates degree to which the precipitation is peaked near mean value.



High kurtosis

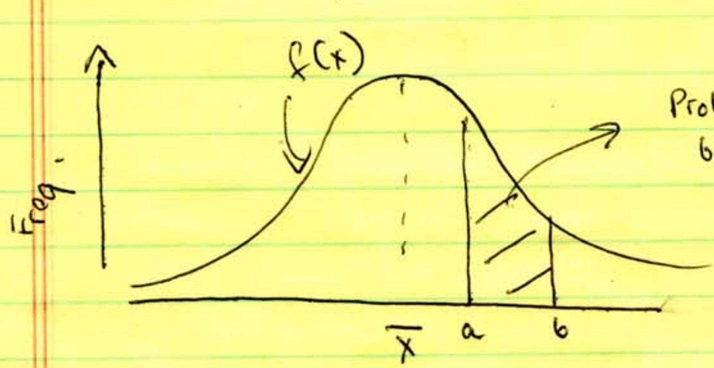


Low kurtosis.

Normal distrubation : kurtosis = 3.

Probability distributions

Can express the probability that a selected variable falls between a and b on a frequency distribution plot, or histogram.



Probability of lying between a & b.

$$= \int_a^b f(x)\, dx.$$

f(x) is referred to as the probability density function.

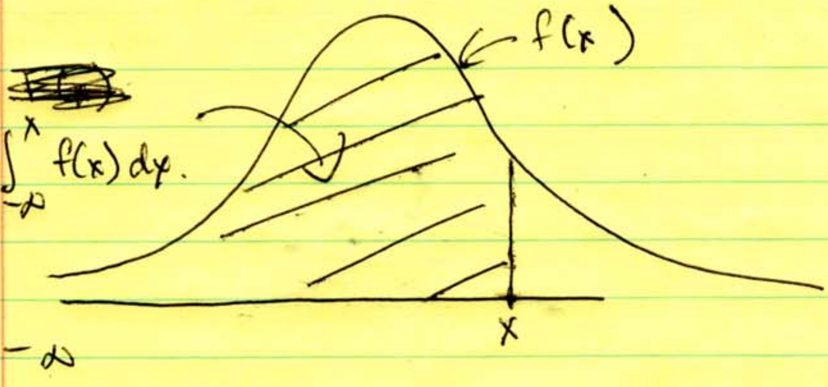
Considering the whole range of values, the probability must be equal to 1.

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

Cumulative distribution function (F(x)) can be defined as the probability that a variable assumes a value less than x.

$$F(x) = \int_{-\infty}^{x} f(x)\, dx.$$

$$F(x) = \int_{-\infty}^{x} f(x)\,dx.$$



most of the time)

The most common distribution which fits geophysical data (and course grades) is the normal or Gaussian distribution.
~~Most people know this More commonly~~
More commonly known as the "bell curve"
→ Focus on this first, then move to non-normal dist.

<u>Normal distribution.</u>

For μ = population mean
σ = standard deviation.

The "Bell curve" or normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\}$$

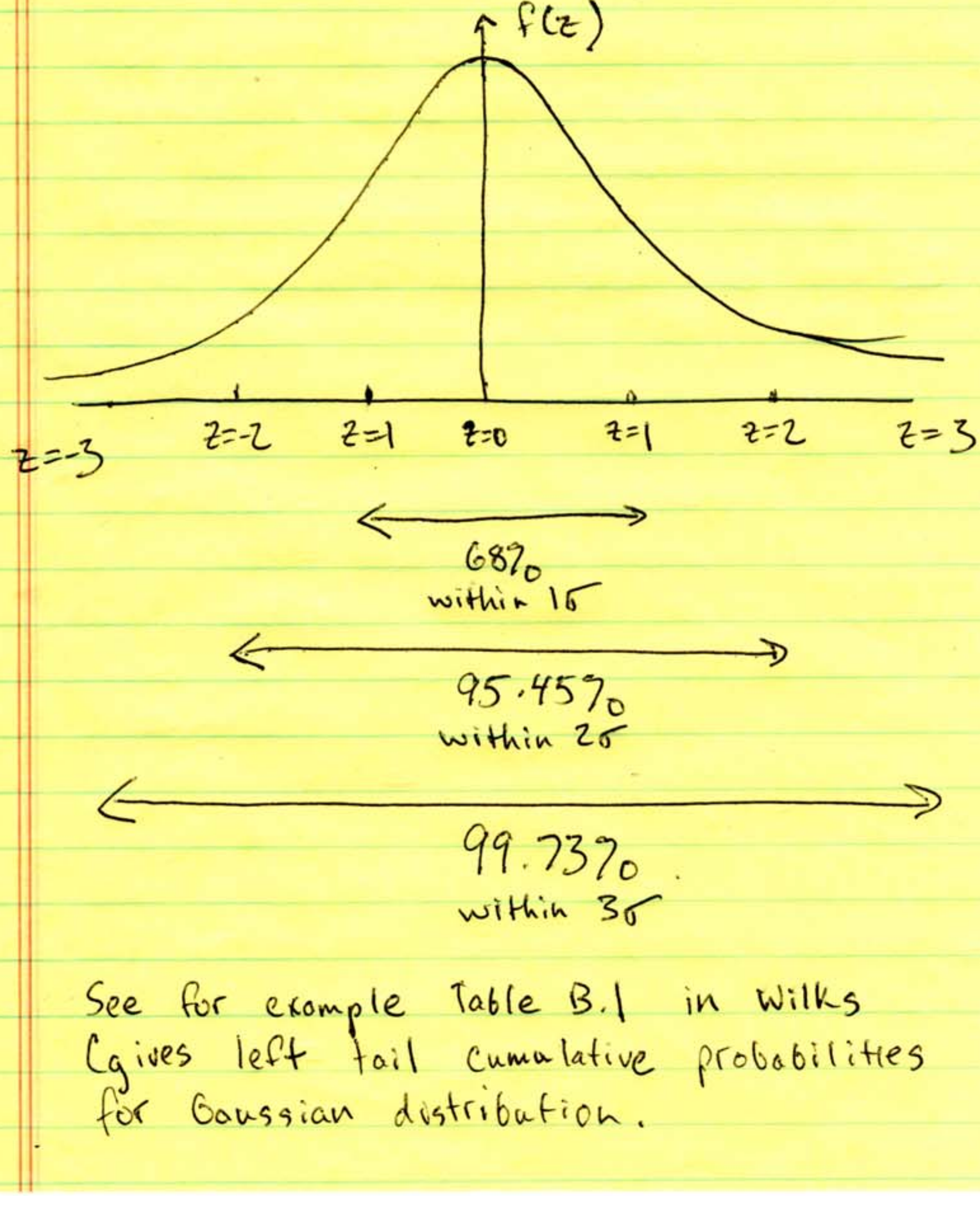
If we consider a normalized variable z

$$z = \frac{x-\mu}{\sigma}$$

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} z^2\right\}$$

Cumulative probability distribution:

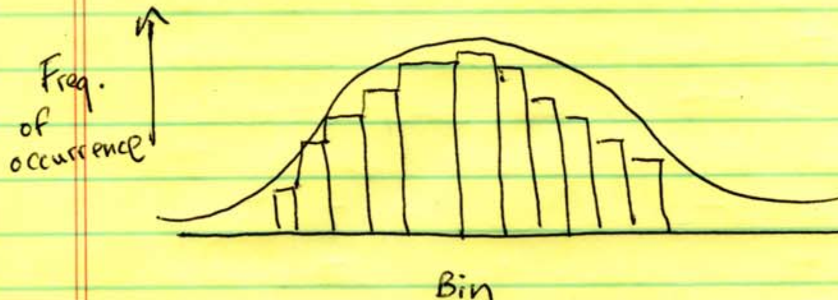
$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} z^2\right\} dz.$$



See for example Table B.1 in Wilks (gives left tail cumulative probabilities for Gaussian distribution.

Lect. 3

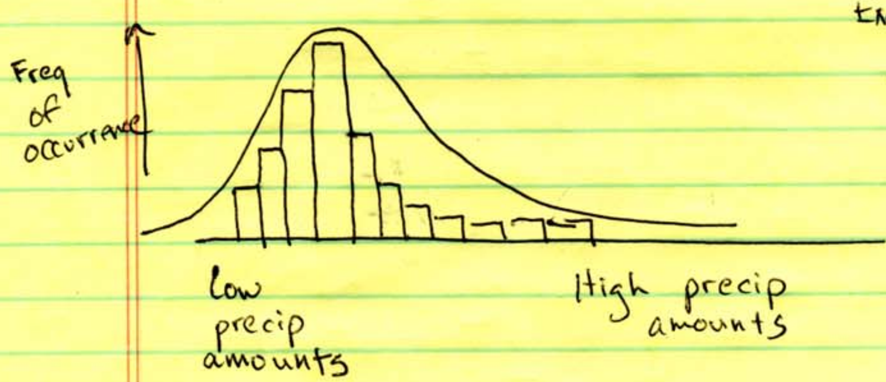
Most geophysical fields are normally distributed, but it is always a good idea to plot a histogram along with the normal distribution just to make sure it is a good assumption for whatever dataset you're working with..



Example of a normally distributed histogram.

Some examples where it may not work: precipitation, land surface variables like soil moisture and vegetation.

Example of a positively skewed histogram.



There are goodness of fit tests that can be applied to verify a given distribution fits whatever data you may have (e.g. Gaussian, gamma, etc.) More on this later...

Data are typically normally distributed (i.e. follow a bell curve) if the sample size is large ($N > 30$)

The mean of an infinite sample size is denoted by μ. The sample mean ($\bar{x}$) is computed from a finite sample size N.

The standard deviation of the sampling distribution of means, or <u>standard</u> <u>error</u> <u>estimate</u> of the mean is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

σ = std. deviation of the population.

Function of the standard deviation <u>and</u> the sample size:

As the sample size decreases, there is a greater uncertainty in the estimate of the mean.