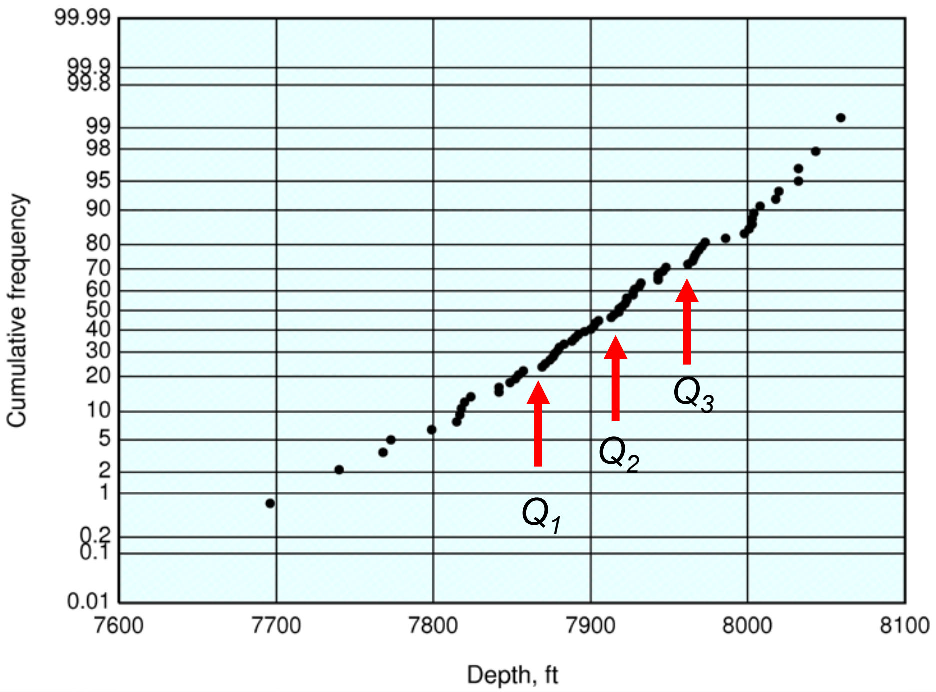
UNCF QUARTILES

Q_1 = 7,871.75 ft

Q_2 = 7,918 ft

Q_3 = 7,965.75 ft



Q_2 coincides with the median.

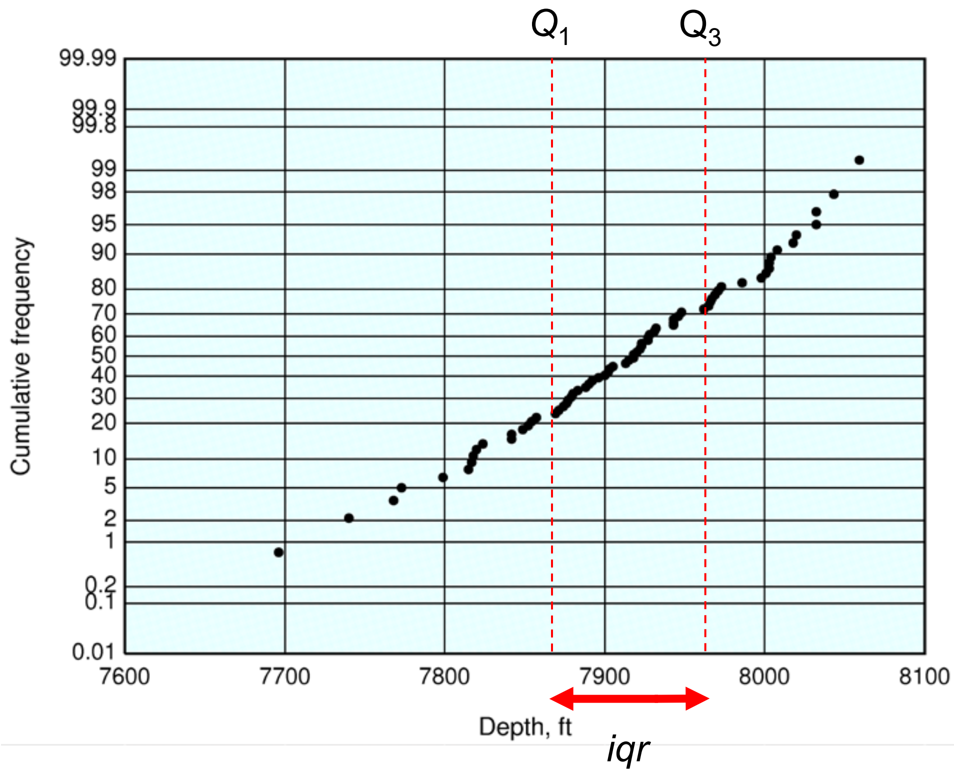
INTERQUARTILE RANGE

The interquartile range, *iqr*, is the difference between the upper and the lower quartiles

$$iqr = Q_3 - Q_1,$$

thus measuring the central spread of the data.

For the UNCF sample, the interquartile range is 94 ft.

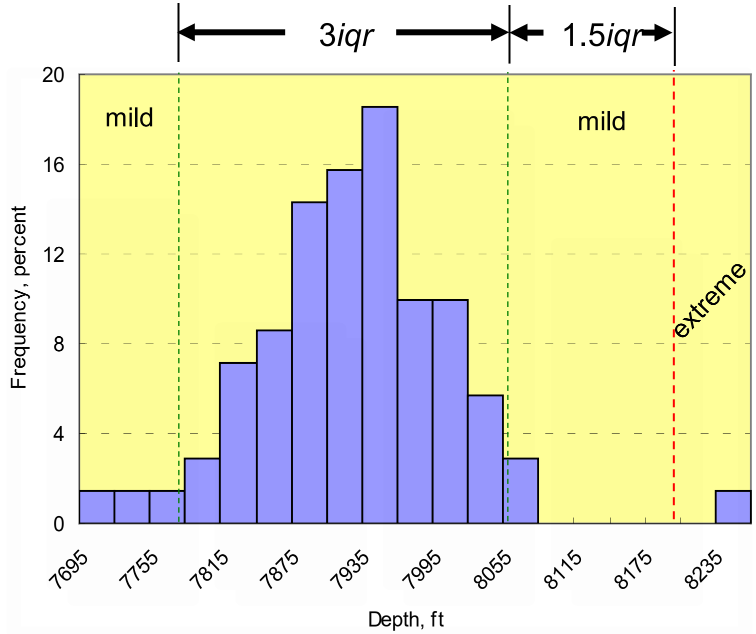


The interquantile range is more robust than the variance but insensitive to values in the lower and upper tails.

OUTLIER

Outliers are values so markedly different from the rest of the sample that they raise the suspicion that they may be from a different population or that they may be in error, doubts that frequently are hard to clarify. In any sample, outliers are always few, if any.

A practical rule of thumb is to regard any value deviating more than 1.5 times the interquartile range, *iqr*, from the median as a mild outlier and a value departing more than 3 times *iqr* as an extreme outlier.



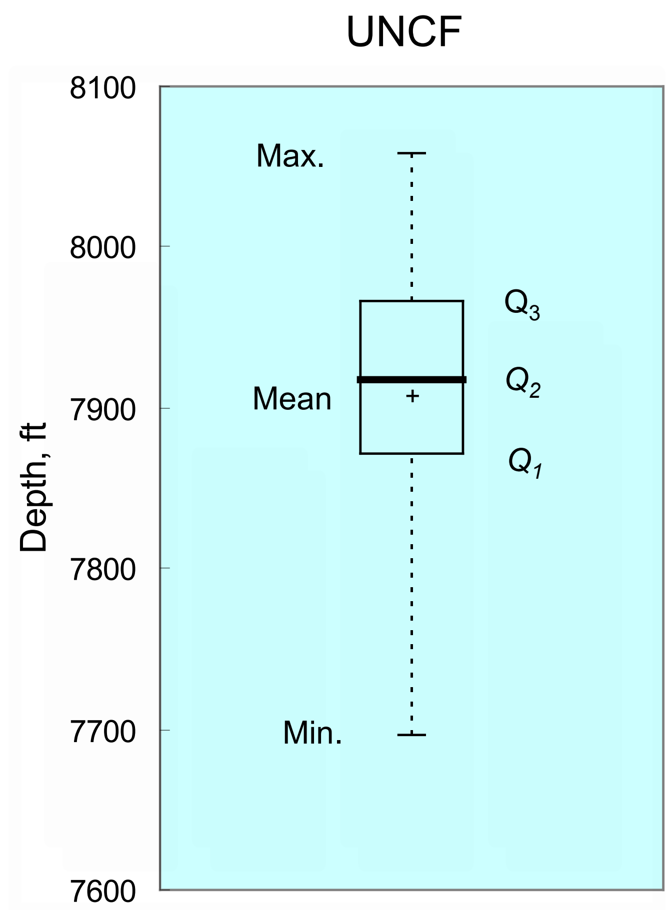
For the UNCF sample, all mild outliers seem to be legitimate values, while the extreme outlier of 8,240 ft is an error.

BOX-AND-WHISKER PLOT

The box-and-whisker plot is a simple graphical way to summarize several of the statistics:

- Minimum
- Quartiles
- Maximum
- Mean

Variations in this presentation style abound. Extremes may exclude outliers, in which case the outliers are individually plotted as open circles. Extremes sometimes are replaced by the 5th and 95th percentiles.



USGS

MEASURES OF SHAPE

The most commonly used measures of shape in the distribution of values are:

- Coefficient of skewness
- Quartile skew coefficient
- Coefficient of kurtosis

COEFFICIENT OF SKEWNESS

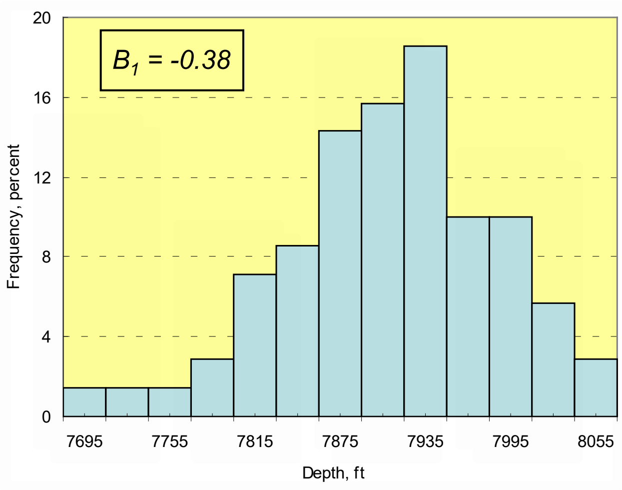
The coefficient of skewness is a measure of asymmetry of the histogram. It is given by:

$$B_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - m)^3}{\sigma^3}$$

If : $B_1 < 0$, the left tail is longer;

$B_1 = 0$, the distribution is symmetric;

$B_1 > 0$, the right tail is longer.

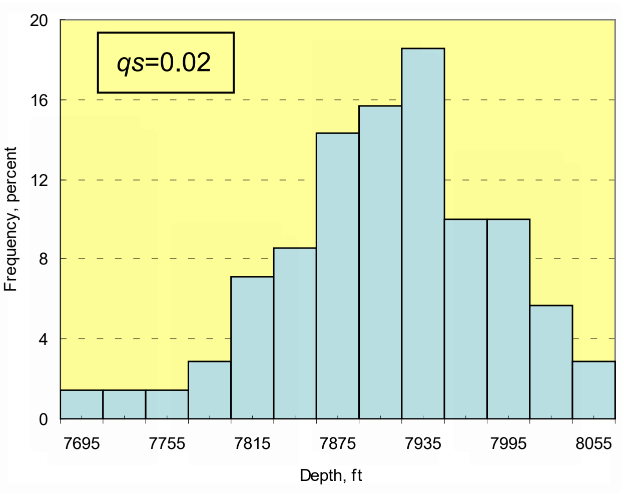


The UNCF coefficient of skewness is -0.38.

QUARTILE SKEW COEFFICIENT

The quartile skew coefficient serves the same purpose as the coefficient of skewness, but it is more robust, yet only sensitive to the central part of the distribution. Its definition is:

$$qs = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{iqr}$$



If : $qs < 0$, the left tail is longer;

$qs = 0$, the distribution is symmetric;

$qs > 0$, the right tail is longer.

The UNCF quartile skew coefficient is 0.02.

COEFFICIENT OF KURTOSIS

This statistic measures the concentration of values around the mean. Its definition is:

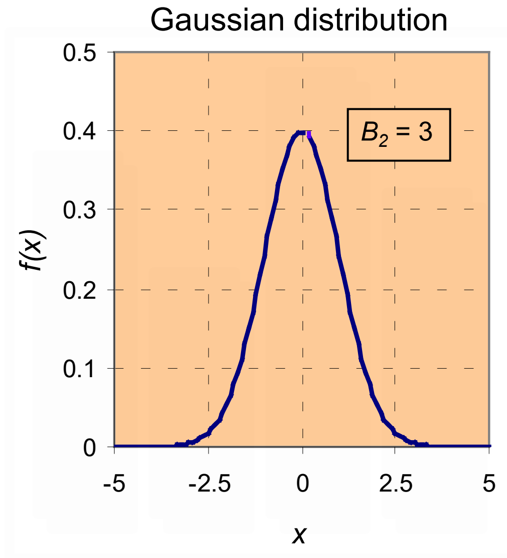
$$B_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - m)^4}{\sigma^4}$$

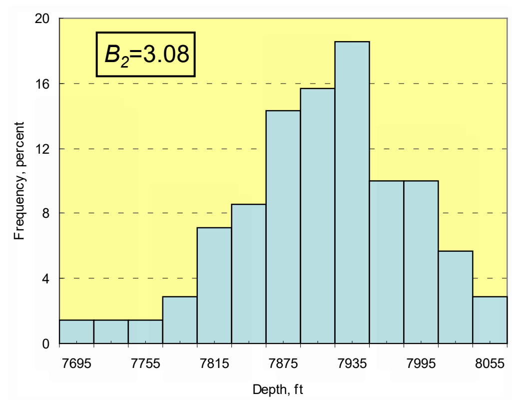
If : $B_2 < 3$, the distribution is more peaked than the Gaussian distribution;

$B_2 = 3$, it is as peaked as the Gaussian;

$B_2 > 3$, it is less peaked than the Gaussian.

The UNCF coefficient of kurtosis is 3.08.





USGS

MODELS

PROBABILITY

Probability is a measure of the likelihood that an event, *A*, may occur. It is commonly denoted by Pr[*A*].

- The probability of an impossible event is zero, Pr[*A*] = 0. It is the lowest possible probability.
- The maximum probability is Pr[*A*] = 1, which denotes certainty.
- When two events *A* and *B* cannot take place simultaneously, Pr[*A* or *B*] = Pr[*A*] + Pr[*B*].
- Frequentists claim that Pr[*A*] = N_A/N, where N is total number of outcomes and N_A the number of outcomes of *A.* The outcomes can be counted theoretically or experimentally.
- For others, a probability is a degree of belief in *A*, even if no random process is involved nor a count is possible.

JOINT AND CONDITIONAL PROBABILITIES

Let A and B be two random events. The probability of their joint (simultaneous) occurrence, $\Pr[A \text{ and } B]$, is

$$\boxed{\Pr[A \text{ and } B] = \Pr[A \cap B] = \Pr[A \mid B] \cdot \Pr[B]}$$

where $\Pr[A \mid B]$ is the conditional probability of A given that B has occurred.

If $\Pr[A \mid B] \neq \Pr[A]$, then the result of one event affects the probability of the other. Such types of outcomes are said to be statistically dependent. Otherwise, they are statistically independent. For example:

- Heads and tails in successive flips of a fair coin are independent events.
- Being dealt a king from a deck of cards and having two kings on the table are dependent events.

BAYES'S THEOREM

This is a widely used relationship for the calculation of conditional probabilities. Given outcomes *A* and *B*, because $\Pr[A \cap B] = \Pr[B \cap A]$:

$$\boxed{\Pr[A \mid B] = \frac{\Pr[B \mid A]}{\Pr[B]} \Pr[A]}$$

Example

Suppose there are two boxes. A blue ball is drawn (event *B*). What is the probability the ball came from box #1 (event *A*)?

Box	Number of balls		
	Blue	Red	Total
#1	20	5	25
#2	12	18	30
	32	23	55

- If one only knows that there are 2 boxes, Pr[*A*] = ½ = 0.5.
- Now, if the table is available, Pr[*B*|*A*] = 20/25 = 0.8.
- Pr[*B*] = 32/55 = 0.59. Hence:

$$\Pr[A \mid B] = \frac{0.8}{0.59} 0.5 = 0.69$$

PROBABILITY FUNCTIONS

Analytical functions approximating experimental fluctuations are the alternative to numerical descriptors and measures. They provide approximations of general conditions. Their drawback is the loss of fine detail in favor of simpler models.

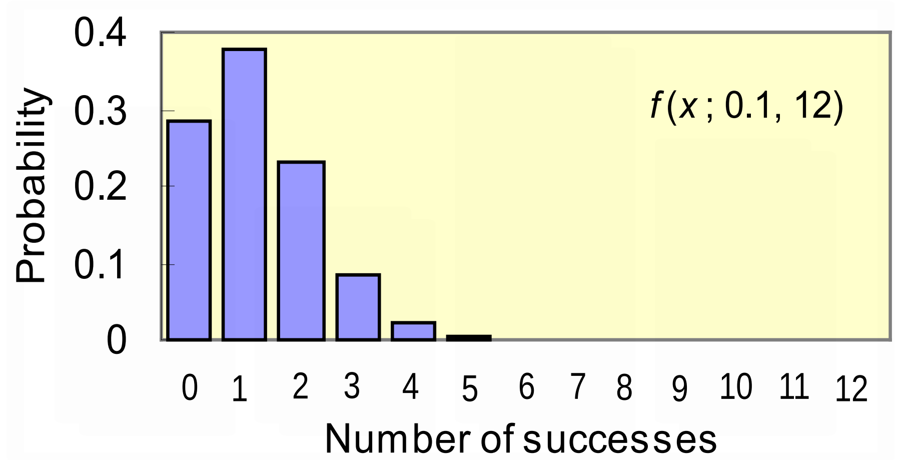
Models approximating histograms are called probability density functions.

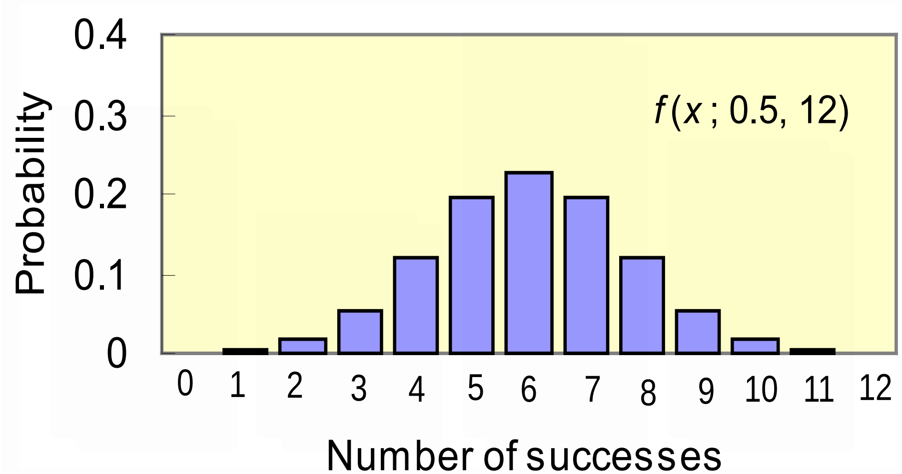
Variations in the parameters of a probability density function allow the generation of a family of distributions, sometimes with radically different shapes.

The main subdivision is into discrete and continuous distributions, of which the binomial and normal distribution, respectively, are typical and common examples.

BINOMIAL DISTRIBUTION

This is the discrete probability density function, $f(x;p,n)$, of the number of successes in a series of independent (Bernoulli) trials, such as heads or tails in coin flipping. If the probability of success at every trial is p, the probability of x successes in n independent trials is

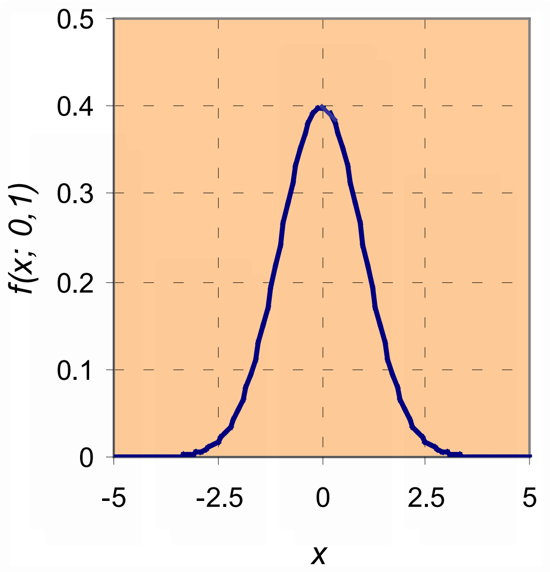




$$f(x;p,n) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \; x = 0, 1, 2, \ldots, n$$

USGS

NORMAL DISTRIBUTION



The most versatile of all continuous models is the normal distribution, also known as the Gaussian distribution. Its parameters are μ and σ, which coincide with the mean and the standard deviation.

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

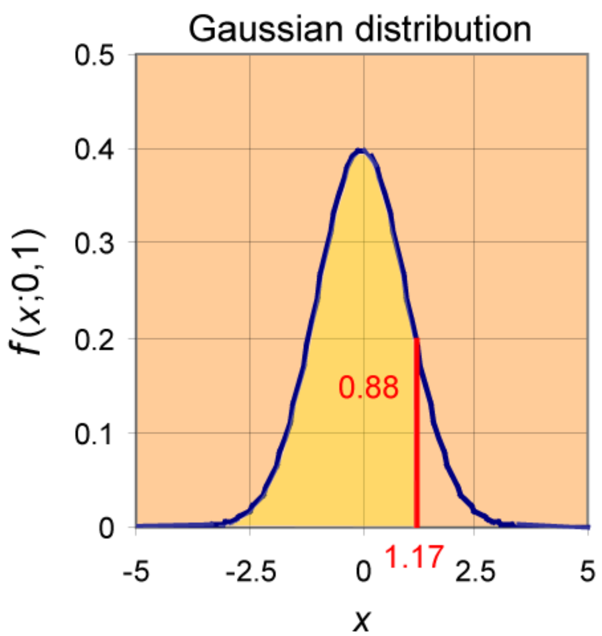
If $X = \log(Y)$ is normally distributed, Y is said to follow a lognormal distribution. Lognormal distributions are positively defined and positively skewed.

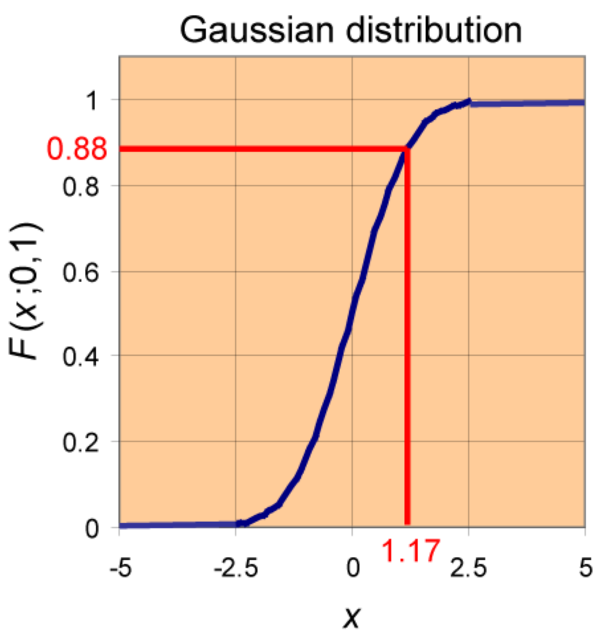
USGS

PROBABILITY FROM MODELS

$$\text{Prob}\left[X \leq x_1\right] = \int_{-\infty}^{x_1} f(x)\,dx \qquad\qquad \text{Prob}\left[X \leq x_1\right] = F(x_1)$$

Examples:





$$\text{Prob}\left[X \leq 1.17\right] = \int_{-\infty}^{1.17} \text{Normal}(x;0,1)\,dx = 0.88$$

$$\text{Prob}\left[X \leq 1.17\right] = F(1.17) = 0.88$$

EXPECTED VALUE

Let X be a random variable having a probability distribution $f(x)$ and let $u(x)$ be a function of x. The expected value of $u(x)$ is denoted by the operator $\mathrm{E}[u(x)]$ and it is the probability weighted average value of $u(x)$.

If X is continuous, such as temperature:

$$\mathrm{E}[u(x)] = \int_{-\infty}^{\infty} u(x) f(x) dx,$$

and if it is discrete, like in coin flipping:

$$\mathrm{E}[u(x)] = \sum_{x} u(x) f(x).$$

In the latter case, for the trivial example of $u(x) = x$, if all values are equally probable, the expected value turns into

$$\mathrm{E}[x] = \frac{1}{n} \sum_{x} x,$$

which is exactly the definition of the mean.

USGS

MOMENT

Moment is the name given to the expected value when the function of the random variable, if it exists, takes the form $(x-a)^k$, where k is an integer larger than zero, called the order.

If a is the mean, then the moment is a central moment.

The central moment of order 2 is the variance. For an equally probable discrete case,

$$M_2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

USGS