

PDF: A *Probability Density function* is a function,  $f(x)$ , whose integral over its domain is equal to 1. Note that if the function is a discrete function, the integral becomes a sum.

CDF: A *Cumulative Distribution Function*,  $F(x)$ , is the integral of a PDF,

$$F(x) = \int_{-\infty}^x f(t) dt.$$

Note that if the function is a discrete function, the integral becomes a sum.

The  $\Gamma$ -function is the function defined by  $\Gamma(n) = \int_0^{\infty} x^{n-1} e^{-x} dx$ .

It has the properties  $\Gamma(n+1) = n\Gamma(n)$ , for positive integers  $n$ ,  $\Gamma(n) = (n-1)!$ , and,  $\Gamma(1/2) = \sqrt{\pi}$ .

The *Gaussian* or it Bell Curve, or *Normal curve* has PDF  $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$ .

The *t-distribution* has its PDF given by

$$t(x, n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n} \Gamma(\frac{n}{2})} \cdot (1 + t^2/n)^{-\frac{n+1}{2}}, \text{ for } -\infty < t < \infty.$$

$n$  is the number of *degrees of freedom* of the distribution.

As  $n \rightarrow \infty$ , the *t-distribution* approaches the Standard Normal, the Gaussian with mean 0 and variance 1.

The *F-distribution* has as its PDF given by

$$F(x, n_1, n_2) = \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \left(\frac{n_1}{n_2}\right)^{n_1/2} \cdot x^{n_1/2-1} \left(1 + \frac{n_1}{n_2}x\right)^{-\frac{1}{2}(n_1+n_2)} \text{ for } x > 0,$$

where  $n_1, n_2$  are the degrees of freedom.

The  $\chi^2$ -distribution has PDF given by  $\chi^2(x, n) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{\frac{n-2}{2}} e^{-x/2}$  for  $x \geq 0$ .

### Distribution expected value, variance

The expected value of a random variable is  $E(X) = \int_{-\infty}^{\infty} xf(x) dx$  if the PDF is continuous,  $\sum_X Xf(X)$  if the PDF is discrete.

The variance of a random variable is  $V(X) = E(X^2) - E(X)^2$ .

The standard deviation of a random variable is the square root of its variance.

### Sample mean, variance

The sample mean of a statistical sample,  $x_1, x_2, \dots, x_n$ , is the average value of the samples,  $\mu = \sum_{i=1}^n x_i/n$ .

The sample variance of a statistical sample is given by  $\sigma^2 = \sum_{i=1}^n (x_i - \mu)^2/(n-1)$ .

The sample standard deviation is the square root of the sample variance.

### Binomial Distribution

The binomial distribution is followed when two outcomes occur (e.g. “heads”(H) or “Tails”(T), “win” or “loss”). Suppose that one of the events, e.g. H, happens with probability  $p$ , where  $p$  is constant over all trials.

Then after  $n$  independent trials, the probability that there will be  $k$  instances of H is

$$\binom{n}{k} p^k (1-p)^{n-k}.$$

Note that the sum of these over  $k$  from 0 to  $n$  is 1.

*Proof:*  $\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1^n = 1.$

The expected number of H is  $\sum_{k=0}^n k \cdot \binom{n}{k} p^k (1-p)^{n-k}$ . Since  $k \cdot \binom{n}{k} = \frac{n! \cdot k}{k!(n-k)!} = n \cdot \frac{(n-1)!}{(k-1)!(n-1-(k-1))!}$ , the expected number of H is  $\sum_{k=0}^n n \cdot \binom{n-1}{k-1} p \cdot p^{k-1} (1-p)^{n-k} = np \cdot \sum_{k=0}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = np \cdot 1 = np.$

The variance is found by looking at  $E(X^2) = \sum_{k=0}^n k^2 \cdot \binom{n}{k} p^k (1-p)^{n-k}$ .

$$\begin{aligned} E(X^2) &= \sum_{k=0}^n k^2 \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^n k \cdot np \cdot \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\ &= \sum_{k=0}^n (1 + (k-1)) \cdot np \cdot \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\ &= np + np \cdot \sum_{k=0}^n (k-1) \cdot \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\ &= np + np \cdot (n-1)p = np + n(n-1)p^2 \end{aligned}$$

So the variance is  $E(X^2) - E(X)^2 = np + n(n-1)p^2 - n^2p^2 = np - np^2 = np(1-p).$

### Estimation of difference between means

We wish to estimate the difference between two mean values taken from normally distributed data. Suppose that one sample, of size  $n_1$ , has mean  $X_1$  and variance  $\sigma_1^2$ , while another sample, of size  $n_2$ , has mean  $X_2$  and variance  $\sigma_2^2$ . We would like to determine the difference between the means of the samples.

The difference will be normally distributed with mean  $X_1 - X_2$  and variance  $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ .

### Estimation of the ratio of two variances

If  $S_1^2$  and  $S_2^2$  are the variances of independent random samples of sizes  $n_1$  and  $n_2$ , respectively, taken from normal populations, then  $F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$  follows an  $f$ -distribution with  $n_1 - 1$  and  $n_2 - 1$  degrees of freedom.

We may then estimate

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{f_{\alpha/2, n_1-1, n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot f_{\alpha/2, n_2-1, n_1-1}.$$

### Testing a single sample variance

If we wish to determine whether a sample variance  $s^2$  is equal to a particular value  $\sigma_0^2$  we test the value

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

which follows a  $\chi^2$ -distribution with  $n-1$  degrees of freedom.

### Correlation

Two variables are correlated if changes in one affect the other. We may compute a correlation coefficient,  $r$  by

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

It is always the case that  $-1 \leq r \leq 1$ .

If the variables are perfectly correlated, e.g.  $y = 3x + 7$ , then  $r = 1$ .

If the variables are perfectly negatively correlated, e.g.  $y = 10 - 2x$ , then  $r = -1$ .

If there is no relation between the variables, then  $r = 0$ .