ERG 2012B
Advanced Engineering Mathematics II

Part IV
Introduction to Probability & Statistics

Lectures #22
Probability & Statistics Basics

Random Variable

Definition: A **random variable** X is a function with the following properties:

1. X is defined on the sample space S of an experiment and its values are real numbers

2. For every real number a the probability $P(X = a)$ that X takes the value a in a trial is well defined; likewise, for every interval I the probability $P(X \in I)$ that X takes **any** value in I in a trial is well defined.

These probabilities form the **probability distribution** of X given the **distribution function** (or **cumulative distribution function**)

$$F(x) = P(X \leq x)$$

the probability that X takes **any** value not exceeding x

The discrete distribution is given by the **probability function** of X defined by:

$$f(x) = \begin{cases} p_j & \text{if } x = x_j \quad (j = 1,2,\ldots) \\ 0 & \text{otherwise} \end{cases}$$

Random Variable

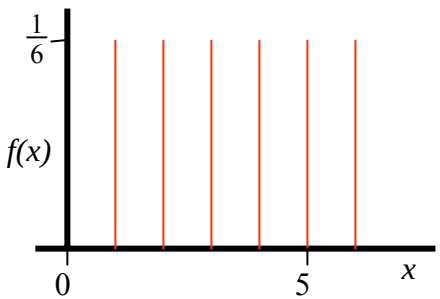
From this we get the values of the distribution function $F(x)$ by taking sums:

$$F(x) = \sum_{x_j \le x} f(x_j) = \sum_{x_j \le x} p_j$$

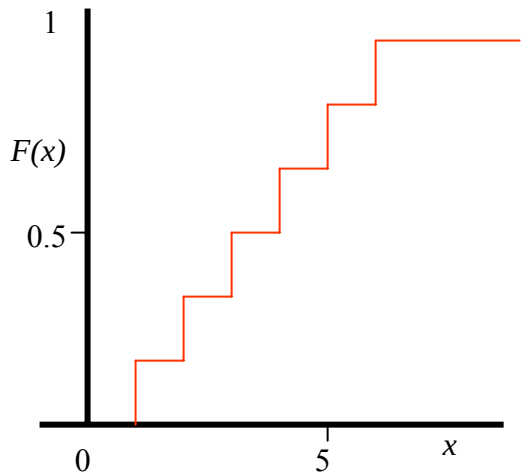
where for any given x we sum all the probabilities p_j for which x_j is smaller than or equal to x. This is a **step function** with upward jumps of size p_j at the possible values of x_j of X and constant in between

Example: The probability function $f(x)$ and the distribution function $F(x)$ of the discrete random variable:

X = Number a fair dice turns up



X has possible values x=1,2,3,4,5,6 each with probability 1/6



Probability & Distribution Functions

Example: The random variable

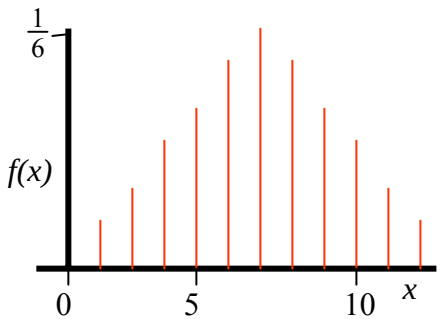
X = Sum of the two Numbers two fair dice turn up

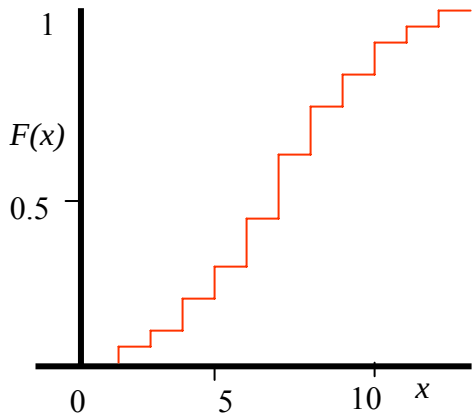
is discrete and has possible values 2, 3,4,….,12. There are 36 equally likely outcomes:

(1,1), (1,2),…, (6,6)

Now X=2 occurs in the case (1,1); X=3 twice: (1,2) and (2,1); X=4 thrice: (1,3), (2,2), (3,1), etc.. Hence $f(x)=P(X=x)$ and $F(x)=P(X\leq x)$ have values:

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36
$F(x)$	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	36/36





Examples

In the previous example compute the probability of a sum of at least 4 and at most 8.

Solution: $P(3 < X \leq 8) = F(8) - F(3) = 26/36 - 3/36 = 23/36$

Waiting problem. Countably infinite sample space. In tossing a fair coin, let X = *Number of trials until the first head appears.*

Then:

$P(X = 1) = P(H) = 1/2$ (H=head)

$P(X = 2) = P(TH) = 1/2 \cdot 1/2 = 1/4$ (T=tail)

$P(X = 3) = P(TTH) = 1/2 \cdot 1/2 \cdot 1/2 = 1/8\ldots$ etc.

and in general $P(X = n) = (1/2)^n$, $n = 1,2,\ldots.$

Mean

The **mean value** or **mean** of a distribution is denoted by μ and is defined by:

$$\mu = \sum_j x_j f(x_j) \qquad \text{(discrete distribution)}$$

$$\mu = \int_{-\infty}^{\infty} x f(x) dx \qquad \text{(continuous distribution)}$$

In the first expression $f(x)$ is the probability function of the random variable X considered and we sum over all possible values. By definition it is assumed that the sum converges

In the second $f(x)$ is the density of X. By definition it is assumed that the integral exists.

A distribution is said to be **symmetric** wrt a number $x=c$ if for every real x $\quad f(c+x) = f(c-x)$

Mean & Variance

Theorem: Mean of a symmetric Distribution

If a distribution is symmetric with respect to $x=c$ and has a mean μ then $\mu=c$

The **variance** of a distribution is denoted by σ^2 and is defined by the formula:

$$\sigma^2 = \sum_j (x_j - \mu)^2 f(x_j) \qquad \text{(discrete distribution)}$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \qquad \text{(continuous distribution)}$$

By definition it is assumed that the series converges and the integral exists.

For a discrete distribution with $f(x)=1$ at a point and $f(x)=0$ otherwise, we have $\sigma^2 =0$, otherwise $\sigma^2 > 0$

The +ve square root of the variance is called the **standard deviation**. Both are a measure of the spread of a distribution

Example 1

Mean and Variance

The random variable

X = *Number of heads in a single toss of a fair coin*

Has the possible values $X = 0$ and $X = 1$
with probabilities $P(X=0) = ½$ and $P(X=1) = ½$

From the definition of the mean we have:

$$\mu = 0.\ ½ + 1.½ = ½$$

And from the definition of the variance we have:

$$\sigma^2 = (0 - ½)^2.½ + (1 - ½)^2.½ = ¼$$

Example 2

Uniform Distribution

The distribution with the density

$$f(x) = 1/(b - a) \qquad \text{if } a < x < b$$

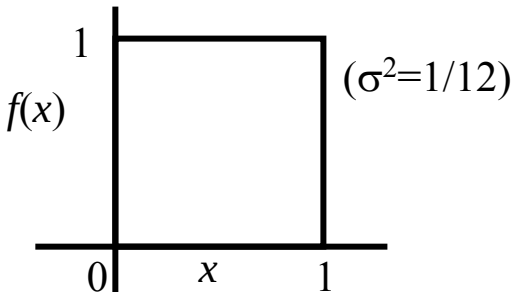
and $f = 0$ otherwise is called a **uniform distribution** on the interval $a < x < b$.

From the definition of the mean we have:

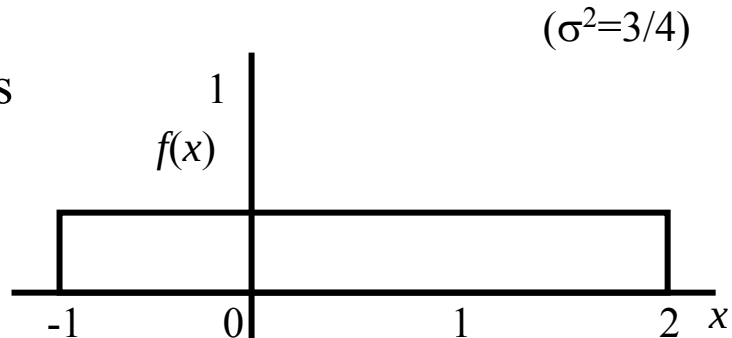
$$\mu = \int_a^b \frac{x}{b-a}\,dx = \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$$

And from the definition of the variance we have:

$$\sigma^2 = \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{1}{b-a}\,dx = \frac{(b-a)^2}{12}$$



Note: these distributions have the same mean but different variances and larger variance gives larger spread



Expectation, Moments

For any random variable X and any continuous function $g(X)$ defined for all real X, the **mathematical expectation** of $g(X)$ is defined by

$$E(g(X)) = \sum_j g(x_j) f(x_j) \qquad \text{(discrete distribution)}$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx \qquad \text{(continuous distribution)}$$

where f is the probability function and the density of X respectively

Note: for $g(X) = X$ this gives the mean of X i.e. $\mu = E(X)$

In general taking $g(X) = X^k$ (k=1,2,..) we get the ***k*th moment** of X given respectively by

$$E(X^k) = \sum_j x_j^{\,k} f(x_j) \quad \text{and} \quad E(X^k) = \int_{-\infty}^{\infty} x^k f(x) dx$$

Central Moments

Taking $g(X) = (X - \mu)^k$ gives the ***k*th central moment**

$$E((X-\mu)^k) = \sum_j (x_j - \mu)^k f(x_j) \quad \text{and} \quad \int_{-\infty}^{\infty} (x-\mu)^k f(x)dx$$

Notice that the 2nd central moment (k=2) is the variance:

$$\sigma^2 = E((X - \mu)^2)$$

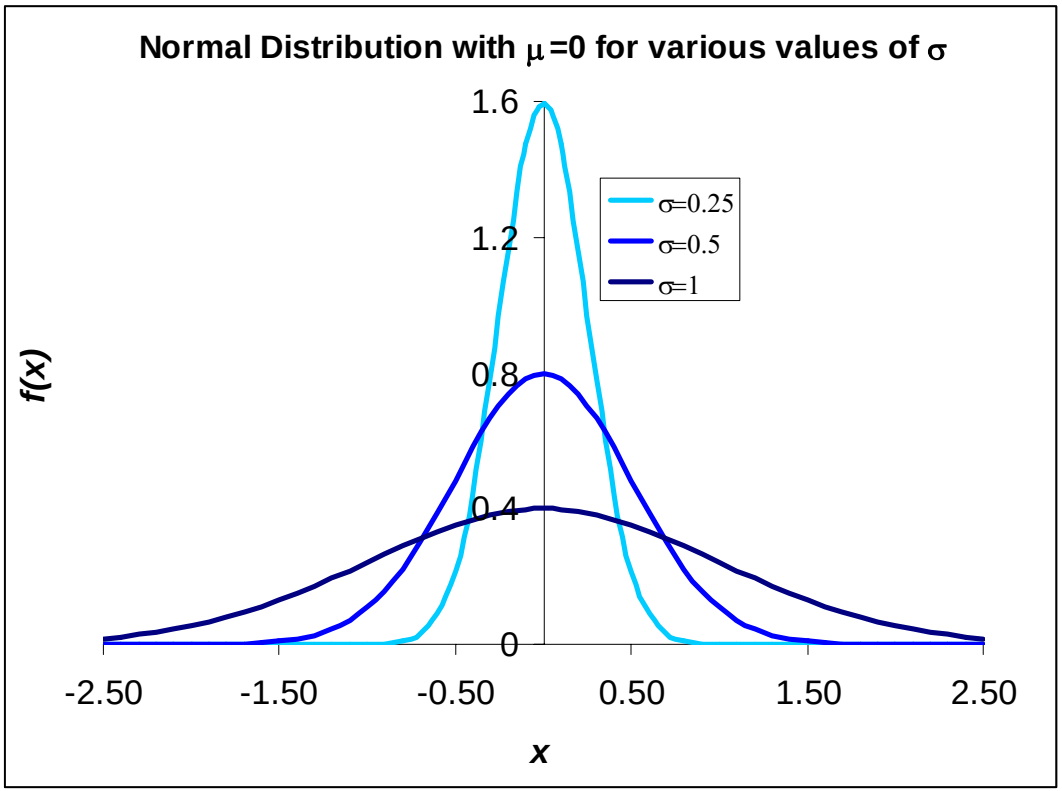
and

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

Normal Distribution

The **normal** or **Gauss distribution** is defined with the density:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (\sigma > 0)$$



The density curves are *bell-shaped* and have a peak at $x=\mu$.

σ^2 is the variance and we see that for small σ^2 we have a high peak and steep slopes and as σ^2 increases the density spreads out.

Normal Distribution

The **distribution function** $F(x)$ is obtained from the density function:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(v-\mu)^2}{2\sigma^2}} dv \qquad (*)$$

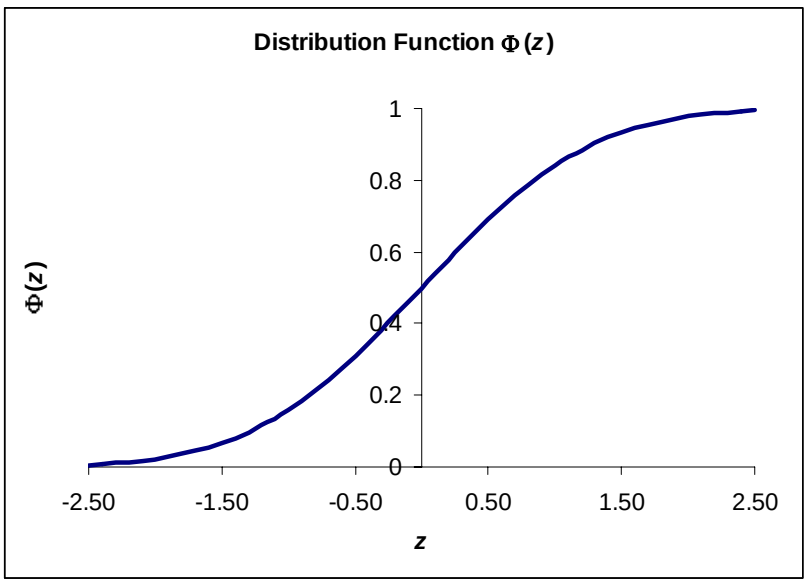
Hence the probability that a normal random variable X assumes any value in some interval $a < x \leq b$ is:

$$P(a < X \leq b) = F(b) - F(a) = \frac{1}{\sigma\sqrt{2\pi}} \int_{a}^{b} e^{-\frac{(v-\mu)^2}{2\sigma^2}} dv$$

Normal Distribution

The integral (*) can not be integrated by calculus but has been tabulated. This is impractical for every μ and σ.
Fortunately, it is enough to do so for the standardized normal random variable $Z = (X - \mu)/\sigma$ with mean 0 and variance 1

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}} du$$



Values of the integral are given in the appendix of the text book.

To get $F(x)$ in terms of $\Phi(z)$ we use the substitution:
$u=(v-\mu)/\sigma$ then $dv = \sigma du$
and the integral becomes:

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-\frac{\mu^2}{2}} dv$$

$$\text{or } F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Normal Distribution

So that the probability that a normal random variable X assumes any value in the interval $a < x \leq b$ is:

$$P(a < X \leq b) = F(b) - F(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

In particular,

if $a = \mu - \sigma$ and $b = \mu + \sigma$ we have $F(x) = \Phi(1) - \Phi(-1)$

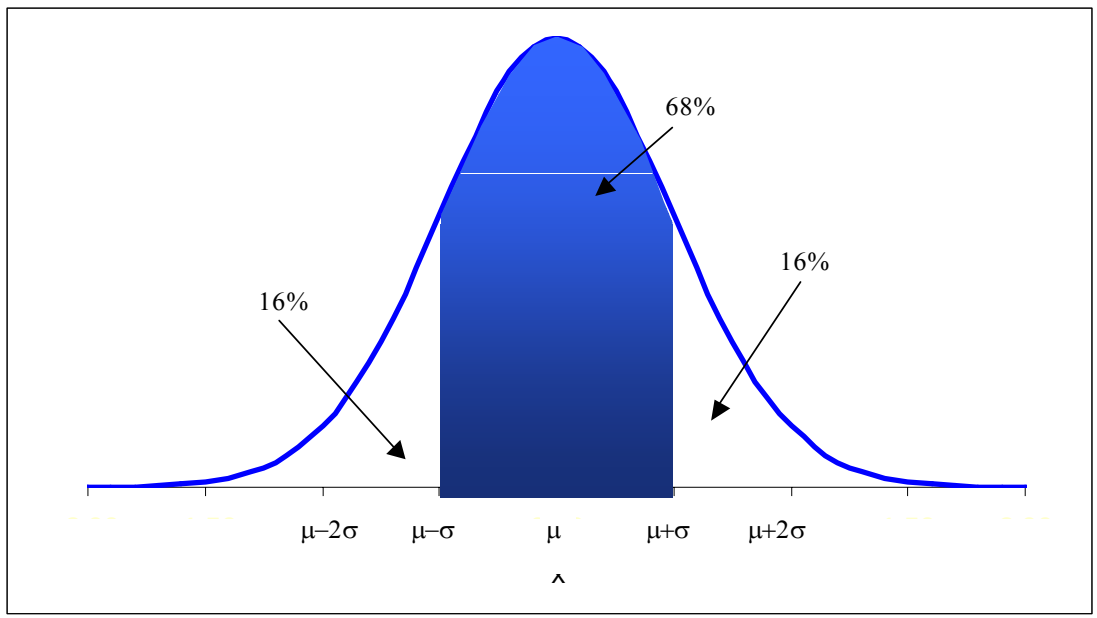
if $a = \mu - 2\sigma$ and $b = \mu + 2\sigma$ we have $F(x) = \Phi(2) - \Phi(-2)$ etc.

From tables we find:

$P(\mu - \sigma < X \leq \mu + \sigma) \approx 0.68$

$P(\mu - 2\sigma < X \leq \mu + 2\sigma) \approx 0.955$

$P(\mu - 3\sigma < X \leq \mu + 3\sigma) \approx 0.997$



Examples

Example 1: For a normal random variable X with mean 0 and variance 1 find the probabilities:

(a) $P(X \le 2.44)$ (b) $P(X \le -1.66)$ (c) $P(X \ge 1)$ (d) $P(2 \le X \le 10)$

Solution: since $\mu=0$ and $\sigma^2=1$ we can get the values directly from the tables:

(a) 0.9927 (b) 0.1230 (c) $1- P(X \le 1) = 1- 0.8413= 0.1587$

(d) $\Phi(10) = 1.0000$, $\Phi(2)= 0.9772$, $\Phi(10) - \Phi(2) = 0.0228$

Example 2: Compute the probabilities above with $\mu=0.8$, $\sigma^2=4$

Solution: from the tables:

(a) $F(2.44) = \Phi((2.44 - 0.8)/2) = \Phi(0.82) = 0.7939$

(b) $F(-1.66) = \Phi(-0.98) = 0.1635$

(c) $1- P(X \le 1) = 1- F(1)= 1- \Phi(0.1) = 0.4602$

(d) $F(10)-F(2) = \Phi(4.6) - \Phi(0.6) = 1 - 0.7257 = 0.2743$

Examples

Example 3: For a normal random variable X with mean 0 and variance 1 determine c such that:

(a) $P(X \geq c) = 0.1$ (b) $P(X \leq c) = 0.05$

(c) $P(0 \leq X \leq c) = 0.45$ (d) $P(-c \leq X \leq c) = 0.99$

Solution: From the tables:

(a) $1-P(X \leq c) = 1- \Phi(c) = 0.1$, $\Phi(c)=0.9$, $c = 1.282$

(b) $c = -1.645$

(c) $\Phi(c) - \Phi(0) = \Phi(c) - 0.5 = 0.45$, $\Phi(c) =0.95$, $c=1.645$

(d) $c = 2.576$

Introduction to Statistics

- In statistics we are concerned with methods for designing and evaluating experiments to obtain information about practical problems that involve processes affected by chance
- The totality of the entities to be studied is called the **population**
- Statistically only a few of these entities – **a sample** – are chosen at **random**, inspected and from the inspection conclusions can be drawn about the whole population.
- Such conclusions are not absolutely certain but we can obtain measures for the reliability of the conclusions obtained from the samples by statistical methods.

Introduction to Statistics

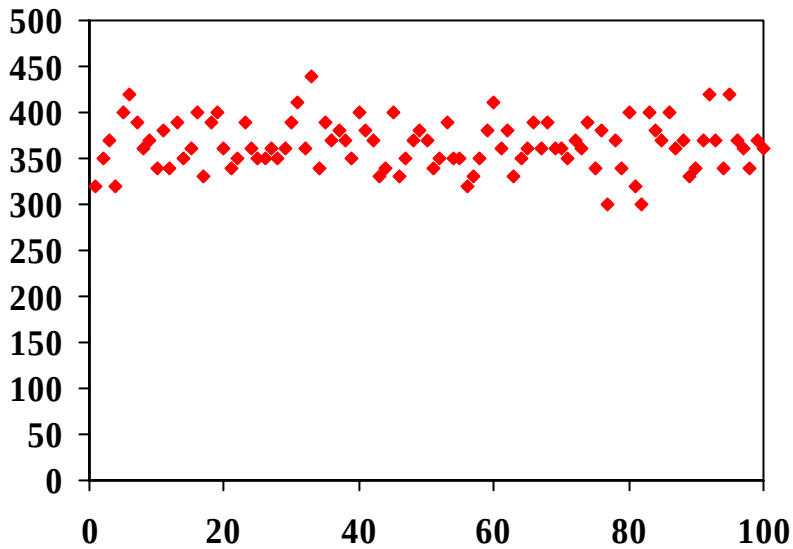
- Problems of differing natures may require different methods, but the steps leading to the formulation and solution of a problem are similar in most cases. They are:
- **Formulation of the problem:** describe the problem in a precise fashion and limit the investigation – need to get a useful answer in a prescribed interval of time, need to ensure all concepts well defined
- **Design of experiment:** the choice of the statistical method to be used, the sample size and the physical methods to be used
- **Data collection:** adhere to the rules decided on above
- **Data processing:** data arranged in clear form, and sample parameters (mean, variance etc.) calculated
- **Statistical inference:** conclusions are drawn from the sample data

Processing of Samples

In the course of a statistical experiment we normally obtain a sequence of observations. These should be recorded in the order in which they occur. They are know as **sample values**. The number of them is the **sample size** *n*.

Example

320	380	340	410	380	340	360	350	320	370
350	340	350	360	370	350	380	370	300	420
370	390	390	440	330	390	330	360	400	370
320	350	360	340	340	350	350	390	380	340
400	360	350	390	400	350	360	340	370	420
420	400	350	370	330	320	390	380	400	370
390	330	360	380	350	330	360	300	360	360
360	390	350	370	370	350	390	370	370	340
370	400	360	350	380	380	360	340	330	370
340	360	390	400	370	410	360	400	340	360



Sample of 100 values of the tensile strength (kg/cm^2) of concrete cylinders

Frequency Distribution

For a given sample we call $\tilde{f}(x)$ the **frequency function** of the sample and say that it determines the **frequency distribution** of the sample.

The relative frequency satisfies $0 \le \tilde{f}(x) \le 1$

and $\sum_x \tilde{f}(x) = 1$

The frequency function $\tilde{f}(x)$ of a sample is an empirical counterpart or analogue of the probability function $f(x)$ of the corresponding population – although these functions are conceptually quite different: most obviously, a population has **one** $f(x)$, but if we take 10 samples from the same population, we will generally get **10 different** sample frequency functions

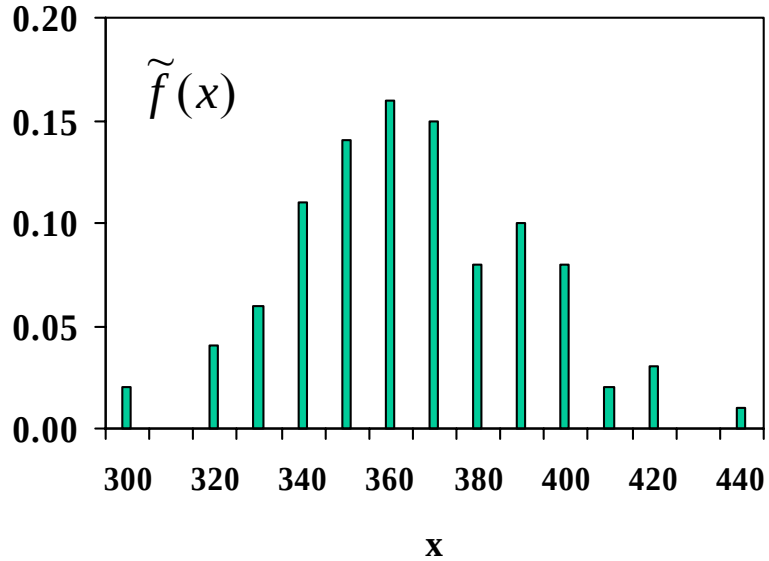
Frequency Distribution

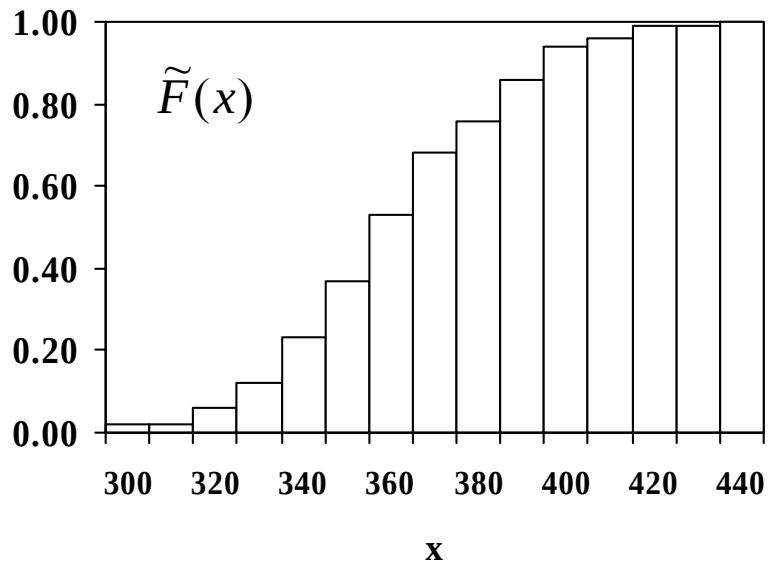
Example

From our previous data we can tabulate the frequency data

The Cumulative frequency function $\widetilde{F}(x)$ is defined similarly to $F(x)$

Tensile Strength x (kg/cm^2)	Absolute Frequency	Relative Frequency	Cumulative Absolute Frequency	Cumulative Relative Frequency
300	2	0.02	2	0.02
310	0	0.00	2	0.02
320	4	0.04	6	0.06
330	6	0.06	12	0.12
340	11	0.11	23	0.23
350	14	0.14	37	0.37
360	16	0.16	53	0.53
370	15	0.15	68	0.68
380	8	0.08	76	0.76
390	10	0.10	86	0.86
400	8	0.08	94	0.94
410	2	0.02	96	0.96
420	3	0.03	99	0.99
430	0	0.00	99	0.99
440	1	0.01	100	1.00





Mean and Variance

The Sample Mean $\bar{x}$ of a sample $x_1, x_2, \ldots x_n$ is defined by

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

The Sample variance s^2 of a sample $x_1, x_2, \ldots x_n$ is defined by

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2)$$

The positive square root of the sample variance is called the **standard deviation** denoted by s

Note: the difference between s^2 and σ^2 – the factor $n/(n\text{-}1)$ - derives from the fact that in calculating s^2 we do not know the value of the true mean μ only an estimate $\bar{x}$. For large n the difference becomes negligible.

see http://mathworld.wolfram.com/Variance.html for a more detailed explanation

Example

Ten randomly selected nails had the lengths (cm):
0.80 0.81 0.81 0.82 0.81 0.82 0.80 0.82 0.81 0.81
Find the mean and variance of the sample

Solution: the mean is simply

$$\bar{x} = \frac{1}{10}(0.80 + 0.81 + 0.81 + ... + 0.81) = 0.811\,\text{cm}$$

The sample variance is given by

$$s^2 = \frac{1}{9}((0.800 - 0.811)^2 + ... + (0.810 - 0.811)^2) = 0.000054\,\text{cm}^2$$

alternatively we can use the frequency data so that:

$$\bar{x} = \frac{1}{10}(2 \cdot 0.80 + 5 \cdot 0.81 + 3 \cdot 0.81) = 0.811\,\text{cm}$$

$$s^2 = \frac{1}{9}(2(0.800 - 0.811)^2 + 5(0.810 - 0.811)^2 + 3(0.820 - 0.811)^2)$$
$$= 0.000054\,\text{cm}^2$$

Estimation of Parameters

Parameters – quantities appearing in distributions, such as p in the binomial distribution and μ and σ in the normal distribution

A point estimate of a parameter is a number computed from a given sample as an approximation of the unknown exact value of the parameter

As an approximation of the mean μ of a population we can take the mean $\bar{x}$ of a corresponding sample. So that the estimate $\hat{\mu}$

$$\hat{\mu} = \bar{x} = \frac{1}{n}(x_1 + x_2 \ldots + x_n)$$

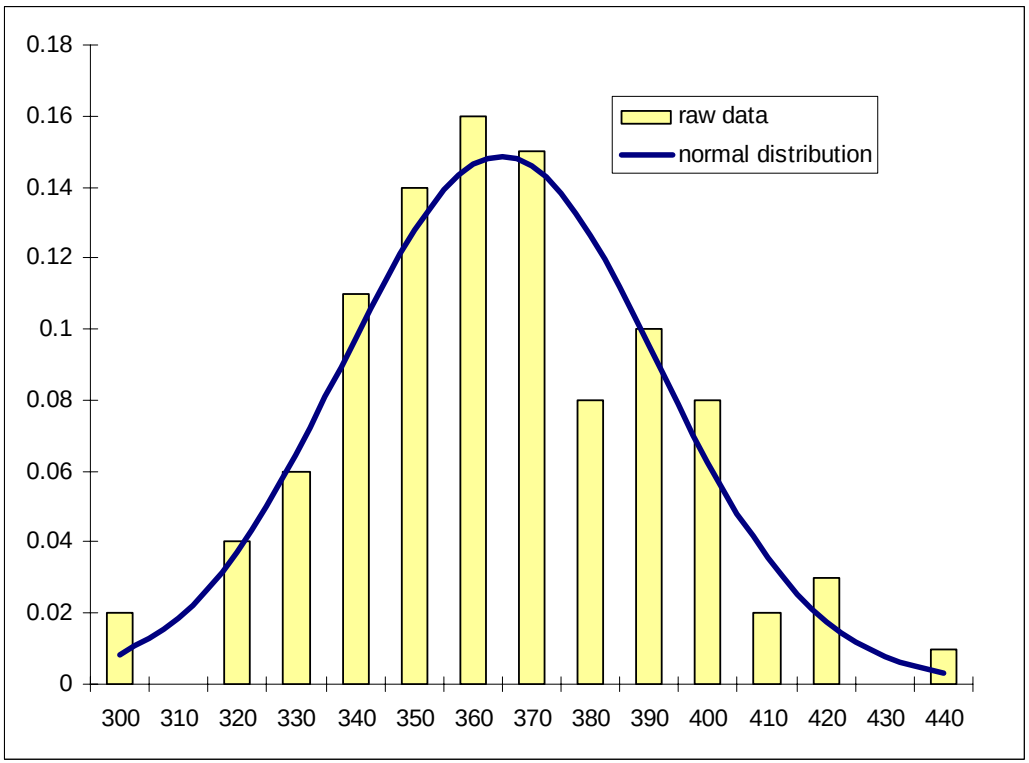
similarly we can estimate the variance of a population from the variance s^2 of a sample.

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

Estimation of Parameters

We can use these estimates to substitute for the real things and obtain estimates for other parameters. For example in the binomial distribution $p=\mu/n$ and so we can make an estimate of p from

$$\hat{p} = \frac{\bar{x}}{n}$$



We can use $\bar{x}$ and s^2 in the normal distribution to provide a fit to our sample data (from the concrete example)

$$\bar{x} = 364.7$$
$$s^2 = 720.1$$