Topics in Probability and Statistics

A Fundamental Construction

Suppose $\{S, P\}$ is a sample space (with probability P), and suppose $X : S \to \mathbb{R}$ is a random variable. The *distribution of* X is the probability P_X on $\mathbb{R}$ defined by

$$P_X(B) = P(X \in B) = P(\{s \in S \mid X(s) \in B\}), \quad \text{for} \quad B \subset \mathbb{R}.$$

We then have a new sample space with probability, namely $\{\mathbb{R}, P_X\}$, and the function x (*i.e.* $f(x) = x$) is a random variable **which has the same distribution as** X:

$$P_x(B) = P_X(x \in B) = P_X(B).$$

A similar construction can be made with several random variables; if X, Y are random variables on S, then the *joint distribution of* X, Y is the probability $P_{x,y}$ on $\mathbb{R}^2$ given by

$$P_{X,Y}(B) = P((X, Y) \in B) = P(\{s \in S \mid (X(s), Y(s)) \in B\}), \quad \text{for} \quad B \subset \mathbb{R}^2.$$

Then $\{\mathbb{R}^2, P_{X,Y}\}$ is a new sample space, and the functions x, y on $\mathbb{R}^2$ are random variables with the same joint distribution as X, Y. This all can be done for any collection of random variables and is called the *product space representation* because the random variables are represented as the coordinate functions on a product of copies of $\mathbb{R}$.

Expectation

The definition of *expectation* or *expected value* of a random variable $X : S, P \to \mathbb{R}$ is supposed to formalize the intuitive idea of the "average value" of X. As we know, the elementary notion of the average of a collection of values is the sum of the values divided by the number of values under consideration; furthermore in calculus one defined the average of a function $f(x)$ on an interval $[a, b]$ to be the integral, $(b-a)^{-1} \int_a^b f(x)\, dx$. (One can prove rather easily that this integral is the limit of the elementary average of the values of f at n equally spaced points in $[a, b]$ as $n \to \infty$.)

It is implicit in the above definitions that the various values under consideration are *weighted* equally in taking the average (*e.g.*, with n values, one weights each with a factor $1/n$.) The expectation of a random variable is quite similar to the average of a function and the general definition involves an integral; the main difference is that one wants to weight the values of X with the probabilities that these values occur.

There are several ways to proceed. One can give a rather *ad hoc* definition of expectation for various types of random variables which is rather elementary, but for which certain

useful properties are not so easy to deduce or understand initially. This is the route taken in our text. On the other hand, one can give a somewhat "fancier" definition for which these useful properties are more transparent, and then observe that in particular cases the fancy definition is just the one given in the text. I will follow this latter course here.

The general definition of the expectation of X, denoted $E(X)$ is

$$E(X) = \int_S X(s)\, P(ds).$$

Of course, one has to define what this means; the idea is that it should be an integral (or sum) of values of X weighted with the probabilities that these values occur. I won't give a complete definition here, but will rather indicate how the definition might be formulated; this will be enough for our purposes.

Suppose first that X is a function which has value 1 on some event A and is equal to 0 elsewhere, *i.e.* $X = 1_A$, the characteristic or indicator function of A. The we define

$$\int_S X(s)\, P(ds) = \int_S 1_A(s)\, P(ds) = P(A).$$

More generally if X is a finite linear combination of characteristic functions, $X = \sum_i c_i \cdot 1_{A_i}$, then

$$\int_S X(s)\, P(ds) = \int_S (\sum_i c_i \cdot 1_{A_i}(s))\, P(ds) = \sum_i c_i \cdot P(A_i).$$

This last formula is true for an infinite sum (series) also, provided the sums involved converge absolutely.

A function of form $X = \sum_i c_i \cdot 1_{A_i}$ is called a *simple* function. We have thus defined the integral of a random variable which is a simple function. For more general random variables X, one proceeds as follows. Find a sequence of random variables X_n which are simple functions and for which $\lim_{n\to\infty} X_n = X$. Then define

$$\int_S X(s)\, P(ds) = \lim_{n\to\infty} \int_S X_n(s)\, P(ds).$$

(Of course, it takes some work to show that this makes sense for a certain class of random variables X, that the definition is independent of the particular sequence chosen, *etc*, *etc*, but this can be done, and it is not necessary to see all the details in order to have an understanding of the resulting ideas.)

An property of the integral which follows easily from the definition is the following: if X and Y are random variables on S

equipped with P, and C and D are constants, then

$$\int_S (C \cdot X(s) + D \cdot Y(s))\, P(ds) = C \cdot \int_S X(s)\, P(ds) + D \cdot \int_S Y(s)\, P(ds).$$

This can also be stated:

$$E(CX + DY) = CE(X) + DE(Y).$$

This is a very important and useful property! (and it is less obvious if one uses the definition of expectation that is given in the text.)

Now suppose that S is a *finite* sample space with P. (The discussion to follow also applies if S is countable, provided the sums discussed converge properly.) Let X be a random variable on S. Then X has only finitely many values, so X is actually a simple function. ($X = \sum_{s \in S} X(s) \cdot 1_{\{s\}}$.) Therefore we have that

$$E(X) = \sum_{s \in S} X(s) P(s).$$

We can often rearrange the above sum to get a possibly more compact expression. Namely, it may be that the set $\{x_1, x_2, ..., x_i, ...\}$ of values of X is considerably smaller than S itself; X may have the same value, x_i, at *many* points of S. Then we could write

$$E(X) = \sum_{s \in S} X(s) P(s) = \sum_i (x_i \cdot \sum_{\{s : X(s) = x_i\}} P(s)) = \sum_i x_i \cdot P(X = x_i).$$

This is the definition given in the text for the expectation of a discrete random variable.

Suppose next that X is a continuous random variable with *pdf* $f(x)$. Assume for simplicity that f is continuous. We can express the expectation of X in this case using $f(x)$ by reasoning as follows. Suppose we pick reals $x_1 < x_2 < ... < x_i < ...$ with $(\Delta x)_i = x_{i+1} - x_i$ small. Then we could approximate X by the simple function which has the values x_i on the set $\{s : x_i \leq X(s) < x_{i+1}\}$ which has probability $\int_{x_i}^{x_{i+1}} f(x)\, dx \approx f(x_i)(\Delta x)_i$. The integral of this simple function is

$$\sum_i x_i \int_{x_i}^{x_{i+1}} f(x)\, dx \approx \sum_i x_i f(x_i)(\Delta x)_i.$$

These latter sums are an approximation to the integral $\int_{-\infty}^{\infty} x f(x)\, dx$, so we see that in this case

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

This is the definition given in the text for the expectation of a continuous random variable.

The formulas derived above indicate that the expectation, $E(X)$, of a random variable X depends only on the distribution of X, and this is indeed true (and not difficult to show, although we omit the proof.) Now we know that X on $\{S, P\}$ has the same distribution as x (the identity function) on $\{\mathbb{R}, P_X\}$. Thus, these have the same expectation:

$$E(X) = \int_S X(s)\, P(ds) = \int_{-\infty}^{\infty} x\, P_X(dx).$$

(The last integral is just a special case of the integral of a random variable discussed above.) We can view this as a general formula including the special cases discussed above for discrete random variables or continuous random variables with a *pdf*. In fact, if $P_X = \sum_i p(x_i)\delta_{x_i}$ (which is the case if X is discrete), then

$$\int_{-\infty}^{\infty} x\, P_X(dx) = \sum_i x_i p(x_i) = \sum_i x_i \cdot P(X = x_i);$$

and if $P_X(dx) = f(x)dx$, then

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

Another notation which is commonly used is to write $dF_X(x)$ for $P_X(dx)$. Here $F_X(x) = P(X \le x) = \int_{-\infty}^{x} P_X(ds)$. For this reason $P_X(dx)$ is considered to be a "generalized" derivative of F_X. (Note that if X has a *pdf* $f(x)$, then $F_X(x) = \int_{-\infty}^{x} f(s)ds$, so $dF_X(x)/dx = f(x)$, or $dF_X(x) = f(x)dx$ in the usual sense if f is continuous.) Thus we have

$$E(X) = \int_S X(s)\, P(ds) = \int_{-\infty}^{\infty} x\, P_X(dx) = \int_{-\infty}^{\infty} x\, dF_X(x).$$

Another fact of considerable importance and usefulness is that if $h(X)$ is a funtion of X, then the expectation of $h(X)$ can be computed using the distribution of X. Precisely,

$$E(h(X)) = \int_{-\infty}^{\infty} h(x)\, P_X(dx) = \int_{-\infty}^{\infty} h(x)\, dF_X(x).$$

For the special cases that X is discrete or continuous with *pdf* f, this amounts to the formulas

$$E(h(X)) = \sum_i h(x_i)P(X = x_i) \quad \text{or} \quad E(h(X)) = \int_{-\infty}^{\infty} h(x)f(x)\, dx.$$

This is often quite convenient, because computing the distribution of $h(X)$ can be quite complicated. However, this not necessary if one just wants $E(h(X))$.

Samples and Sampling Distributions

In practice, what we often have to deal with is a set of numbers $x_1, \ldots, x_n$, *i.e.*, a *sample*. In many cases, we interpret the x_i as values of a collection of independent, identically distributed random variables $X_1, \ldots, X_n$. (The X_i all have a common distribution P_{X_i} and distribution function $F(x)$. Independence of the X_i means that the joint distribution of the X_i is the function $F(x_1, \ldots, x_n) = F(x_1) \cdot \ldots \cdot F(x_n)$. The distribution function $F(x)$ may be unknown or partially known.) A real function $g(x_1, \ldots, x_n)$ of the sample values is called a *characteristic* of the sample. This is a numerical value which we may use to gain information concerning the distribution of the X_i. For example, we could consider the sample mean,

$$g(x_1, \ldots, x_n) = \overline{x} = \frac{x_1 + \ldots + x_n}{n}$$

and try to use this as an estimate of the expectation $E(X_i)$ (which is independent of i.) In order to analyze the usefulness of such a characteristic, we consider also the *random variable* $g(X_1, \ldots, X_n)$. The distribution of this random variable is called the *sampling distribution* of the characteristic g. Another characteristic often used is the sample variance

$$g(x_1, \ldots, x_n) = s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

(Sometimes this is defined with a factor $\frac{1}{n-1}$ instead of $\frac{1}{n}$.)
When we consider such characteristics as random variables, it is probably best to use a capital letter, *e.g.* $\overline{X}$, or S^2, but this is not always done - often, the same symbol is used for the characteristic considered as a numerical quantity or a random variable. For a random variable, we may compute an expectation, variance, *etc*, so, for example, it makes sense to compute the variance of the sample variance: $V(S^2)$. It probably takes some thought to get used to such notions; there are several levels of abstraction involved.

One might reasonably ask whether the sample mean and sample variance (as numerical quantities) are the mean and variance of some random variable. The answer is yes (although they are certainly not usually equal to the mean and variance of the underlying X_i.) To see this, we do the following: corresponding to the sample $x_1, \ldots, x_n$, construct a distribution on the real numbers $\mathbb{R}$, which assigns probability $1/n$ to each of the values $x_1, \ldots, x_n$ (with the understanding that if some value x_i occurs k times among the sample values then the corresponding probability assigned to this value is k/n.) In other words, the probability distribution of the sample is

$$P^* = \frac{1}{n}(\delta_{x_1} + \ldots + \delta_{x_n})$$

(where δ_{x_i} is the discrete probability measure assigning probability 1 to the point x_i and 0 to every other point.) Then $\{\mathbb{R}, P^*\}$ is a sample space, and the function x on this space

is a random variable which we might denote by the symbol X^*. This random variable has the distribution P^*; we denote the distribution function of X^* by $F^*(x)$.
(The function F^* is quite simple:

$$F^*(x) = \frac{1}{n}\{\text{the number of } x_i \leq x\}.)$$

Now once we have a random variable, X^*, we can consider its expectation, variance, moments, *etc.* These are exactly the sample mean, sample variance, sample moments, *etc.* For example, $V(X^*)$ is exactly the sample variance s^2 defined above. (This is one reason for utilizing the factor $\frac{1}{n}$ instead of $\frac{1}{n-1}$, although using the factor $\frac{1}{n-1}$ produces a characteristic whose expectation when considered as a random variable is $V(X_i)$.)

Now $V(X^*)$ and $V(X_i)$ are quite different quantities in principle. However, we want to use $V(X^*) = s^2$ (or perhaps $\frac{n}{n-1}s^2$) as an *estimate* of $V(X_i)$. To decide how good an estimate this is, we need to know something about the distribution of the random variable S^2, *e.g.* $E(S^2)$ and $V(S^2)$. For example, we hope $E(S^2)$ is close to $V(X_i)$, and $V(S^2)$ is small; in such a situation we may consider s^2 as a "good" estimate of $V(X_i)$. One can show (although we don't do the calculation here) that $E(\frac{n}{n-1}S^2) = V(X_i)$, and also that $V(S^2) = O(\frac{1}{n})$ provided that the X_i have finite 4th moments.

We remark that the probability distribution P^* of a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where the $\mathbf{x}_i \in \mathbb{R}^k$, is defined in exactly the same way as in the case of scalar x_i above. The sample distribution function $F^*(\mathbf{x})$ is also defined analogously (replace x_i and x in the above definition by $\mathbf{x}_i$ and $\mathbf{x}$ and interpret the $\mathbf{x}_i \leq \mathbf{x}$ to mean that each component of $\mathbf{x}_i$ is $\leq$ the corresponding component of $\mathbf{x}$.)

The Space of Random Variables with Finite 2nd Moment

Consider all random variables on $\{S, P\}$ with $E(X^2) < \infty$. We also call such random variables *square integrable* since $E(X^2) = \int_S X^2\, P(ds) = \int_{-\infty}^{\infty} x^2\, P_X(dx)$. This is a vector space since $E((X+Y)^2) \leq 2(E(X^2) + E(Y^2))$. (Note that $(x+y)^2 \leq 2(x^2+y^2)$ holds for real x, y.) If X and Y have finite 2nd moments, then $E(XY) < \infty$ (since $|XY| \leq X^2 + Y^2$.) Hence we can define an *inner product* on the square integrable random variables by:

$$< X, Y >= E(XY) = \int_S X(s)Y(s)\, P(ds) = \int\int_{\mathbb{R}^2} xy\, P_{X,Y}(dx\, dy).$$

($P_{X,Y}$ is the joint distribution of X, Y.)
Note that this is the analog of the "dot" product on $\mathbb{R}^n$. For vectors $\mathbf{a}$, $\mathbf{b}$ in $\mathbb{R}^n$,
$< \mathbf{a}, \mathbf{b} >= \mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^n a_i b_i$, a sum of products of components of $\mathbf{a}$ and $\mathbf{b}$; the expectation $E(XY)$ is an integral of products of values of X and Y. Once we have an inner product,

we can do geometry. X and Y are defined to be *orthogonal* if $< X, Y >= E(XY) = 0$, the *length* or *norm* of X is defined to be

$$\|X\| = \sqrt{< X, X >} = \sqrt{E(X^2)}$$

and the *distance* between X and Y is defined to be

$$d(X, Y) = \|X - Y\| = \sqrt{E((X - Y)^2)}.$$

Note that the variance of X is

$$V(X) = \|X - E(X)\|^2$$

the standard deviation of x is

$$\sigma(X) = \|X - E(X)\| = d(X, E(X))$$

and the *covariance* of X and Y is

$$\text{cov}(X, Y) =< X - E(X), Y - E(Y) >= E[(X - E(X))(Y - E(Y))].$$

More generally, if $X_1, \ldots, X_n$ are random variables, their *covariance matrix* is the matrix Λ whose ij entry is

$$\lambda_{ij} = \text{cov}(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))].$$

An important inequality (the *Schwarz Inequality*) is the following:

$$|E(XY)| \ \leq \ \sqrt{E(X^2)E(Y^2)}.$$

This is proved by observing that $E((Y + tX)^2) = E(Y^2) + 2tE(XY) + t^2E(X^2) \geq 0$ for real t; hence the *discriminant* of this quadratic polynomial,
$-4E(X^2)E(Y^2) + 4(E(XY))^2$, must be ≤ 0 (the graph of the quadratic lies in the upper half plane, so there is one or no real roots.) Note that if we replace X and Y by $X - E(X)$ and $Y - E(Y)$, then the inequality becomes

$$|\text{cov}(X, Y)| \ \leq \ \sigma(X)\sigma(Y).$$

From the proof of the Schwarz inequality, it is clear that equality holds exactly when the polynomial $E((Y + tX)^2) = E(Y^2) + 2tE(XY) + t^2E(X^2)$ has a real root (of multiplicity 2 necessarily); in this case, if the root occurs for $t = -C$, then $Y = CX$ (with probability

1.) The same argument shows that if $|\text{cov}(X,Y)| = \sigma(X)\sigma(Y)$, then $Y - E(Y) = C(X - E(X))$, which we can write as $Y = CX + D$ where C, D are constants.

The *correlation coefficient*, $\rho(X,Y)$, of X and Y is defined to be the covariance divided by the product of the standard deviations:

$$\rho(X,Y) = E[(X - E(X))(Y - E(Y))]/\sqrt{E((X - E(X))^2)}\sqrt{E((Y - E(Y))^2)}$$

From the above discussion, we see that $-1 \leq \rho(X,Y) \leq 1$, and if $\rho(X,Y) = \pm 1$, then Y is a linear function of X.

In the section on regression, we will show that $\|Y - (s + tX)\|$ is minimized when

$$t = t_o = \text{cov}(Y,X)/V(X); \qquad s = s_o = E(Y) - E(X)\text{cov}(Y,X)/V(X).$$

The minimum of the square of the distance is then given by

$$\|Y - (s_o + t_o X)\|^2 = V(Y)(1 - \rho^2(X,Y)).$$

Thus, the variance of Y is reduced by a factor of $1 - \rho^2$, and this is the maximun reduction that is possible by subtracting a linear function of X from Y. We can therefore think of the correlation coefficient as a measure of the strength of a linear relation between X and Y.

We denote the collection of square integrable random variables by $\mathcal{L}^2(S,\, P)$. This is a vector space which is complete in the metric $d(X,Y)$ (in the same way that $\mathbb{R}^n$ is complete in the usual metric - *i.e.* Cauchy sequences have limits; if you are not familiar with this notion, that shouldn't affect your understanding of what follows.)

An important notion is the following: If X is in $\mathcal{L}^2(S,\, P)$, and W is a (closed) subspace of $\mathcal{L}^2(S,\, P)$, then the *(orthogonal) projection of X onto W* is the (unique) element $w_o \in W$ which is closest to X, *i.e.* w_o is the unique $w \in W$ minimizing $\|X - w\|$ (or equivalently $\|X - w\|^2$.) Of course, if $X \in W$, then the projection of X on W is X.

(*Note*: There is a "formula" of sorts for the projection of X on W. First construct a basis of the subspace W consisting of mutually orthogonal random variables v, orthogonally project X onto each basis vector using the formula: $\text{proj}(X,v) = < X, v > v/ < v, v >$, and then add up all these projections. This isn't explicit enough to be very useful in general, though. The characterization as the element of W closest to X is often the best way to think about the projection.)

We remark that if we stick with the way we defined the projection of X on W, then it is not necessary that W is a *subspace* of $\mathcal{L}^2(S,\, P)$.

(A subspace is a subset which contains linear combinations of any of its members, ı.e., if $X_1, X_2 \in W$, then $C_1X_1 + C_2X_2 \in W$ also for any constants C_1, C_2.)

The same construction works if W is a (closed) *affine* subset, *i.e.*, a translate of a subspace by a fixed vector (the analog in $\mathbb{R}^3$ is a a line, plane, *etc* which doesn't pass thru the origin), or more generally any nonempty closed convex subset.
("Closed" means that any sequence in W converging to something in the ambient space $\mathcal{L}^2(S,\, P)$, actually converges to something in W. Subspaces of $\mathbb{R}^n$ are closed, but subspaces of infinite domensional spaces are not necessarily closed. A set is convex if for each pair of points in the set, the segment joining them is also in the set.)

The Normal Distribution

A random variable X has a normal distribution if its distribution has a *pdf*

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

This *pdf* is denoted $N(\mu, \sigma^2)$. Then $E(X) = \mu$, and $V(X) = \sigma^2$. Now, suppose $X_1, \ldots, X_n$ are independent, identically distributed $N(0, 1)$ random variables. The joint distribution has *pdf*

$$\frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_i x_i^2}.$$

Suppose $Y_1, \ldots, Y_n$ are random variables with

$$Y_i = \sum_{i=1}^{n} a_{ij} X_j + \mu_i, \quad i = 1, \ldots, n$$

where the a_{ij} and μ_i are constants. Setting $\mathbf{Y}, \mathbf{X}, \mu$ equal to column vectors consisting of the Y_i, X_i, μ_i respectively and A equal to the matrix with entries a_{ij}, these last relations are equivalent to the equation

$$\mathbf{Y} = A\mathbf{X} + \mu.$$

The Y_i are said to have a joint normal distribution. We will determine the joint *pdf* of the Y_i assuming that A is nonsingular (invertible). (If A is singular the distribution is concentrated on a subspace of dimension $< n$ of $\mathbb{R}^n$, but the calculations are similar.) Let R be a subset of $\mathbb{R}^n$. Then

$$P(\mathbf{Y} \in R) = P(A\mathbf{X} + \mu \in R) = P(\mathbf{X} \in A^{-1}(R - \mu))$$
$$= \frac{1}{(2\pi)^{n/2}} \int_{A^{-1}(R-\mu)} e^{-\frac{1}{2}\mathbf{x}\cdot\mathbf{x}}\, dx_1 \ldots dx_n = (\text{setting } \mathbf{x} = A^{-1}(R - \mu))$$
$$\frac{1}{(2\pi)^{n/2}} \int_R e^{-\frac{1}{2}A^{-1}(\mathbf{y}-\mu)\cdot A^{-1}(\mathbf{y}-\mu)} \,|\det A|^{-1}\, dy_1 \ldots dy_n.$$

Now setting $(A^{-1})^t A^{-1} = (AA^t)^{-1} = \Lambda^{-1}$, the last integral above becomes

$$\frac{1}{(2\pi)^{n/2}\sqrt{\det \Lambda}} \int_R e^{-\frac{1}{2}\Lambda^{-1}(\mathbf{y}-\mu)\cdot(\mathbf{y}-\mu)}\, dy_1 \dots dy_n$$

which exhibits the joint *pdf* of the Y_i. The matrix Λ is the covariance matrix of the Y_i, *i.e.*, the ij entry λ_{ij} of Λ is $\mathrm{cov}(Y_i, Y_j)$. To see this note that

$$\mathrm{cov}(Y_i, Y_j) = \mathrm{cov}(\sum_k a_{ik}X_k, \sum_l a_{jl}X_l) = \sum_{k,l} a_{ik}a_{jl}\mathrm{cov}(X_k, X_l) = \sum a_{ik}a_{jk} = (A^tA)_{ij}.$$

(The next to the last equality follows from the fact that $\mathrm{cov}(X_k, X_l) = \delta_{kl}$ (= 1 or 0 according as $k = l$ or $k \neq l$.)
If $\mathrm{cov}(Y_i, Y_j) = 0,\ i \neq j$, then Λ is a diagonal matrix with diagonal entries $\sigma_1^2, \dots, \sigma_n^2$ (with $\sigma_i^2 = V(Y_i)$), and then the joint *pdf* of the Y_i is

$$\frac{1}{(2\pi)^{n/2}\sqrt{\det \Lambda}} \int_R e^{-\frac{1}{2}\sum_i \frac{(y_i-\mu_i)^2}{\sigma_i^2}}\, dy_1 \dots dy_n$$

which factors into a product, and the Y_i are independent. Hence we have the important fact:
If jointly normal random variables are uncorrelated (have mutual covariance 0), then they are independent.

An important result concerning normal random variables is the following:

Theorem: Suppose $X_1, \dots, X_n$ are independent, identically distributed normal random variables with $E(X_i) = \mu, V(X_i) = \sigma^2$.
Let $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ (where $\overline{X} = (X_1 + \dots + X_n)/n$.)
Then $(n-1)S^2/\sigma^2$ has the distribution of a sum of $n-1$ squares of independent, identically distributed $N(0,1)$ random variables (this is called a *chisquare* distribution with $n-1$ *degrees of freedom*), so $E(S^2) = \sigma^2$, and $V(S^2) = 2\sigma^4/(n-1)$.
Furthermore, S^2 and $\overline{X}$ are independent.

Remark: The expression for $E(S^2)$ is correct even if the X_i are not normal (as long as they are independent.) The expression for $V(S^2)$ follows from the known variance of the chisquare distribution, and normality is necessary; however, if the X_i are just independent, but have finite $4th$ moments, then $V(S^2) = O(\frac{1}{n})$, so S^2 is still an unbiased, consistent estimator of σ^2.

Proof of the theorem: Assume first that the X_i are indepenent and identically distributed with mean $\mu = 0$, finite variance, but not necessarily normal. Then

$$\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n (X_i^2 - 2X_i\overline{X} + \overline{X}^2) = \sum_{i=1}^n X_i^2 - n\overline{X}^2.$$

Suppose that A is an *orthogonal* matrix (*i.e.* $AA^t = A^tA = I$) and whose first row is $\frac{1}{\sqrt{n}}, \ldots \frac{1}{\sqrt{n}}$. (It doesn't matter what the other rows are as long A is an orthogonal matrix.) Then

$$A \begin{pmatrix} X_1 \\ \vdots \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} \sqrt{n}\,\overline{X} \\ U_1 \\ \vdots \\ U_{n-1} \end{pmatrix}$$

and because A is orthogonal we have

$$\sum_{i=1}^{n} X_i^2 = n\overline{X}^2 + \sum_{i=1}^{n-1} U_i^2 = n\overline{X}^2 + \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

(The last equality makes use of the equation in the first line of the proof.)
Now the covariance matrices for

$$\begin{pmatrix} X_1 \\ \vdots \\ \vdots \\ X_n \end{pmatrix} \text{ and } \begin{pmatrix} \sqrt{n}\,\overline{X} \\ U_1 \\ \vdots \\ U_{n-1} \end{pmatrix}$$

are equal, since if $\mathbf{Y} = A\mathbf{X}$, where the X_i are independent and identically distributed with mean 0, and A is orthogonal, then

$$\begin{aligned} \operatorname{cov}(Y_i, Y_j) = E(Y_iY_j) = E(\sum_k a_{ik}X_k \sum_l a_{jl}X_l) = E(\sum_{k,l} a_{ik}a_{jl}X_kX_l) \\ = \sum_{k,l} a_{ik}a_{jl}E(X_kX_l) = \sum_k a_{ik}a_{jk}E(X_k^2) = \delta_{ij}E(X_k^2). \end{aligned}$$

Hence

$$E(\sum_{i=1}^{n-1} U_i^2) = E(\sum_{i=2}^{n} X_i^2) = (n-1)E(X_i^2) = E(\sum_{i=1}^{n} (X_i - \overline{X})^2)$$

(The equality of the first and last terms follows from the third equation in the proof.)
Hence

$$E(\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2) = E(X_i^2)$$

and now replacing X_i by $X_i - E(X_i)$ shows that $E(S^2) = \sigma^2$ even if the X_i have nonzero mean. Now suppose that the X_i are normal. Since they are uncorrelated (orthogonal since the means are still assumed $= 0$), so are the $\sqrt{n}\overline{X}, U_1 \ldots, U_{n-1}$, so $\overline{X}$ is independent of the U_i and hence of $\sum_{i=1}^{n-1} U_i^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2$. Also $\sum_{i=1}^{n} (X_i - \overline{X})^2$ has the distribution of

the sum of $n-1$ squares $\sum_{i=1}^{n-1} U_i^2$ of normal independent random variables with $V(U_i) = V(X_i)$, $E(U_i) = 0$. Dividing by σ^2, we get that $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$ has a chisquare distribution with $n-1$ degrees of freedom. If the X_i don't have mean 0, we can replace them by $X_i - E(X_i)$ and argue as above. ∎

The preceeding argument involving orthogonal transformations of normal random variables occurs in many calculations in statistics, and it is worth stating the result in a general form; the idea seems to have been used systematically by R. Fisher. The following is

Fisher's Lemma

Suppose $X_1, \ldots, X_n$ are *iid* $N(0, \sigma^2)$, and suppose $Y_1, \ldots, Y_k$ are defined by

$$Y_i = a_{i1}X_1 + \ldots + a_{in}X_n, \quad i = 1, \ldots, k$$

where the vectors $(a_{i1}, \ldots, a_{in})$ are orthonormal, *i.e.*

$$\sum_{j=1}^{n} a_{ij}a_{lj} = \delta_{il}.$$

Then the quadratic form

$$Q = \sum_{i=1}^{n} X_i^2 - Y_1^2 - \ldots - Y_k^2$$

has the distribution of the sum of $n-k$ independent $N(0, \sigma^2)$ random variables, and Q is independent of $Y_1, \ldots, Y_k$.

To see this, observe that we may find $n-k$ further vectors $(a_{i1}, \ldots, a_{in}), i = k+1, \ldots, n$ so that the matrix with entries a_{ij} is orthogonal, and define

$$Y_i = a_{i1}X_1 + \ldots + a_{in}X_n, \quad i = k+1, \ldots, n.$$

By orthogonality, the Y_i are *iid* $N(0, \sigma^2)$, and

$$\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i^2$$

so

$$Q = \sum_{i=k+1}^{n} Y_i^2.$$

The fact that Q is independent of $Y_1, \ldots, Y_k$ follows from the result proved earlier that jointly normal random variables which are uncorrelated are independent.

Regression

The notion of regression arises in several different ways in statistics. In my opinion, many texts don't provide a good explanation of what is going on. My remarks here are an attempt at an explanation of the various ideas. In the following discussion, all random variables are assumed to have finite 2nd moments, *i.e.*, they are in $\mathcal{L}^2(S, P)$. Suppose X and Y are 2 such random variables. Sometimes we would like to express Y as a function, $f(X)$, of X "as well as possible" (even though it might **not** be the case that $Y = f(X)$.) The solution to this problem has been discussed above; let $\mathcal{F}(X)$ be the subspace of $\mathcal{L}^2(S, P)$ consisting of random variables which are functions of X, and define

$$E(Y|X) = \text{the orthogonal projection of } Y \text{ on } \mathcal{F}(X).$$

In other words, $E(Y|X)$ is the unique function $f(X)$ minimizing

$$d(Y, g(X)) = \|Y - g(X)\| = \sqrt{E((Y - g(X))^2)}$$

as $g(X)$ ranges over all functions of X in $\mathcal{L}^2(S, P)$. If we consider the function $f(x)$ for which this minimum occurs as a function of a real variable x, this function is usually denoted

$$f(x) = E(Y|X = x).$$

(*Note*: In some cases it is possible to define this function $E(Y|X = x)$ directly in terms of the joint distribution of X and Y, and then $E(Y|X)$ is defined composing $E(Y|X = x)$ with X. In fact, this is the way it is done in most statistics texts, if it is done at all. To do this, one has to first define the conditional probability distribution of Y given $X = x$ which is somewhat problematical since often $P(X = x) = 0$, and we sketch this briefly for the case of continuous random variables; the discrete case is similar. Suppose $f(x, y)$ is the joint *pdf* of X, Y, and $f(x)$ is the marginal *pdf* of X. The conditional *pdf* of Y given $X = x$ is defined by

$$f(y|x) = f(x, y)/f(x)$$

for those x for which the denominator $f(x)$ is nonzero. The conditional expectation of Y given $X = x$ is then defined to be

$$E(Y|X = x) = \int_{-\infty}^{\infty} y f(y|x)\, dy.$$

This is a function of x, and if we substitute X for x in this function, the result is $E(Y|X)$. This is sometimes useful, but generally it is best to think of $E(Y|X)$ as it is defined at the beginning of this section.)

Sometimes one wants to express Y "as well as possible" as a special type of function of X, for example as a linear (or perhaps one should say affine) function $\beta_0 + \beta_1 X$. The way to do this is fairly clear; take the (orthogonal) projection of Y onto the subspace of $\mathcal{L}^2(S, P)$ spanned by X and the constants (this subspace is "two dimensional".) In this case, there is a more explicit formula for the result. We are seeking to minimize the function

$$h(\beta_0, \beta_1) = E((Y - \beta_0 - \beta_1 X)^2)$$

with respect to the 2 parameters β_0, β_1. This is a calculus problem; the minimum is found by setting the derivatives of $h(\beta_0, \beta_1)$ with respect to the β_i equal to 0. The equations obtained are

$$\beta_0 + \beta_1 E(X) = E(Y); \qquad \beta_0 E(X) + \beta_1 E(X^2) = E(YX).$$

An elementary calculation shows that the solution of these equations is

$$\beta_1 = \mathrm{cov}(Y, X)/V(X); \qquad \beta_0 = E(Y) - E(X)\mathrm{cov}(Y, X)/V(X).$$

We remark that this procedure generalizes in an obvious way. For example, if we have random variables $Y, X_1, \ldots, X_n$, we can express Y "as well as possible" as an affine expression $\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$ by minimizing $E((Y - \beta_0 - \beta_1 X_1 - \ldots - \beta_n X_n)^2)$. This will require the solution of $n + 1$ linear equations in the $n + 1$ unknowns β_i. In this latter situation, the X_i might be functions of a single X, *e.g.* $X_i = X^i$, in which case we are trying to find a good fit to Y in the form of a polynomial $\beta_0 + \beta_1 X + \ldots + \beta_n X^n$ in X. This still involves the solution of *linear* equations in the β_i.

Now suppose that what one has is a *sample* consisting of n points $(x_1, y_1), \ldots, (x_n, y_n)$. We could try to find the straight line of form $y = \beta_0 + \beta_1 x$ which "best fits" these points by minimizing the sum of squares

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

It is easy to see that this is precisely equivalent to the following: define the *sample distribution* P^* on $\mathbb{R}^2$ which assigns probability $1/n$ to each of the points (x_i, y_i) $i = 1, \ldots n$, and let x, y be the coordinate functions on $\{\mathbb{R}^2, P^*\}$ considered as random variables. Then minimize $E((y - \beta_0 - \beta_1 x)^2)$. The solution of this problem has already been obtained above. In the present situation, this becomes:

$$\beta_1 = \frac{\sum (y_i - \overline{y})(x_i - \overline{x})}{\sum (x_i - \overline{x})^2}; \qquad \beta_0 = \overline{y} - \beta_1 \overline{x}$$

where $\overline{x} = \frac{1}{n}\sum x_i$, *etc.* (Summations are over all relevant values of the index i when this is not indicated explicitly. We assume $\sum(x_i - \overline{x})^2 \neq 0$.)
It seems to be standard notation to denote these values for β_1, β_0 by $\hat{\beta}_1$ and $\hat{\beta}_0$, and we will do so in what follows, *i.e.*

$$\hat{\beta}_1 = \frac{\sum(y_i - \overline{y})(x_i - \overline{x})}{\sum(x_i - \overline{x})^2}; \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \tag{r1}$$

. We also put

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i; \qquad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{r2}$$

We now derive a fundamental identity.

$$E((y - \overline{y})^2) = E((y - \hat{y} + \hat{y} - \overline{y})^2) = E((y - \hat{y})^2) + E((\hat{y} - \overline{y})^2) + 2E((y - \hat{y})(\hat{y} - \overline{y})).$$

The last term on the right hand side in the preceeding equation vanishes because

$$nE((y - \hat{y})(\hat{y} - \overline{y})) = \sum(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum(y_i - \hat{y}_i)\hat{y}_i + \sum(y_i - \hat{y})\overline{y},$$

and the 2 sums on the right of the last equation vanish precisely because

$$\frac{\partial}{\partial\beta_i}E((y - \beta_0 - \beta_1 x)^2)|_{\beta_i = \hat{\beta}_i} = 0.$$

Hence

$$E((y - \overline{y})^2) = E((y - \hat{y})^2) + E((\hat{y} - \overline{y})^2)$$

or equivalently

$$\sum(y_i - \overline{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \overline{y})^2. \tag{r3}$$

In order to attach statistical significance to these results, one considers the y_i and sometimes the x_i as values of random variables, and then can study the sampling distributions of the β_i, *etc.* However, if we replace both the y_i and x_i in $(r1)$ by random variables Y_i and X_i, then the distribution of β_1 is generally complicated to compute. A more convenient assumption is that the y_i are values of random variabvles Y_i, but the x_i are just considered as numerical parameters on which the Y_i depend in some way. Specifically, we shall assume that

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

where the e_i are independent and identically distributed with $E(e_i) = 0, V(e_i) = \sigma^2$.
(So $E(Y_i) = \beta_0 + \beta_1 x_i, \;\; V(Y_i) = \sigma^2$.)

Later we shall also assume that the e_i are normal, but for now we omit this assumption. Substituting Y_i for y_i in the equations $(r1), (r2), (r3)$ above, we have

$$\hat{\beta}_1 = \frac{\sum(Y_i - \overline{Y})(x_i - \overline{x})}{\sum(x_i - \overline{x})^2}; \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x} \tag{$r1'$}$$

.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{$r2'$}$$

$$\sum(Y_i - \overline{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \overline{Y})^2. \tag{$r3'$}$$

(Recall that all sums indicated are over $i = 1, 2, \ldots n$.)
The Y_i are not identically distributed unless $\beta_1 = 0$ ($E(Y_i) = \beta_0 + \beta_1 x_i$), and it will be convenient sometimes to work with the identically distributed random variables

$$K_i = \beta_0 + \beta_1 \overline{x} + e_i = Y_i - \beta_1(x_i - \overline{x}). \tag{$r4$}$$

Note that $\overline{K} = \overline{Y} = \beta_0 + \beta_1 \overline{x} + \overline{e}$ and $E(K_i) = E(\overline{K}) = \beta_0 + \beta_1 \overline{x}, \quad V(K_i) = \sigma^2$.

A nice property of the $\hat{\beta}_i$ is that they are unbiased estimators for the β_i. To see this, first note that

$$\hat{\beta}_1 = \frac{\sum(x_i - \overline{x})(K_i - \overline{K})}{\sum(x_i - \overline{x})^2} + \beta_1.$$

From this, it follows immdiately, that $E(\hat{\beta}_1) = \beta_1$. Also, from the last expression for $\hat{\beta}_1$ we have (noting that we may omit $\overline{K}$ since $\sum(x_i - \overline{x}) = 0$),

$$V(\hat{\beta}_1) = V(\frac{\sum(x_i - \overline{x})K_i}{\sum(x_i - \overline{x})^2}) = \frac{\sigma^2}{\sum(x_i - \overline{x})^2}. \tag{$r5$}$$

Hence,

$$E(\hat{\beta}_1^2) = V(\hat{\beta}_1) + \beta_1^2 = \frac{\sigma^2}{\sum(x_i - \overline{x})^2} + \beta_1^2.$$

Next

$$E(\hat{\beta}_0) = E(\overline{Y} - \hat{\beta}_1 \overline{x}) = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} = \beta_0. \tag{$r6$}$$

We defer the computation of $V(\hat{\beta}_0)$ (the calculation is a bit tedious) in order to exhibit an unbiased estimator for σ^2; this will be quite important when we construct confidence limits for the β_i. Consider the equation $(r3')$. We can't use the term $\sum(Y_i - \overline{Y})^2$ directly to estimate σ^2, because the Y_i have different expectations. However, it turns out that

$$E\sum(Y_i - \hat{Y}_i)^2 = (n-2)\sigma^2 \tag{$r7$}$$

and we prove this now. From $(r1')$, $(r2')$, we have

$$E\sum(\hat{Y}_i - \overline{Y})^2 = E(\hat{\beta}_1^2 \sum(x_i - \overline{x})^2) = \sigma^2 + \beta_1^2 \sum(x_i - \overline{x})^2,$$

and

$$E\sum(Y_i-\overline{Y})^2 = E\sum[K_i-\overline{K}+\beta_1(x_i-\overline{x})]^2$$
$$= E\sum[(K_i-\overline{K})^2+2(K_i-\overline{K})\beta_1(x_i-\overline{x})+\beta_1^2(x_i-\overline{x})^2]$$
$$= E\sum[(K_i-\overline{K})^2+\beta_1^2(x_i-\overline{x})^2] = (n-1)\sigma^2+\beta_1^2\sum(x_i-\overline{x})^2,$$

and these last 2 equations together with $(r3')$ yield $(r7)$.
We define

$$SSE=\sum(Y_i-\hat{Y}_i)^2;\quad S^2=\frac{SSE}{n-2};\quad SSY=\sum(Y_i-\overline{Y}_i)^2 \qquad (r8)$$

so S^2 is an unbiased estimator for σ^2.

Next we calculate $V(\hat{\beta}_0)$. We have

$$\hat{\beta}_0^2=(\overline{Y}-\hat{\beta}_1\overline{x})^2=\overline{Y}^2-2\overline{Y}\hat{\beta}_1\overline{x}+\hat{\beta}_1^2\overline{x}^2$$

so

$$V(\hat{\beta}_0)=E(\overline{Y}^2)-2\overline{x}E(\overline{Y}\hat{\beta}_1)+\overline{x}^2E(\hat{\beta}_1^2)-\beta_0^2,$$

and

$$E(\overline{Y}^2)=(\beta_0+\beta_1\overline{x})^2+\frac{\sigma^2}{n};\quad E(\hat{\beta}_1^2)=V(\hat{\beta}_1)+\beta_1^2=\frac{\sigma^2}{\sum(x_i-\overline{x})^2}+\beta_1^2.$$

Now

$$E(\overline{Y}\hat{\beta}_1)=\frac{\sum(x_i-\overline{x})E(Y_i\overline{Y})}{\sum(x_i-\overline{x})^2},$$

and

$$E(Y_iY_j)=E[(\beta_0+\beta_1x_i+e_i)(\beta_0+\beta_1x_j+e_j)]=(\beta_0+\beta_1x_i)(\beta_0+\beta_1x_j)+\delta_{ij}\sigma^2$$

so

$$E(Y_i\overline{Y})=(\beta_0+\beta_1x_i)(\beta_0+\beta_1\overline{x})+\frac{\sigma^2}{n},$$

and (using the fact that $\sum(x_i-\overline{x})=0$ several times) we get

$$E(\overline{Y}\hat{\beta}_1)=\beta_1(\beta_0+\beta_1^2\overline{x}).$$

Finally, putting the previous equations together, we get

$$V(\hat{\beta}_0)=\frac{\overline{x}^2}{\sum(x_i-\overline{x})^2}\sigma^2+\frac{\sigma^2}{n}=\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{\sum(x_i-\overline{x})^2}\right)=\frac{\sum x_i^2}{n\sum(x_i-\overline{x})^2}\sigma^2.$$

If we put $S_{xx} = \sum (x_i - \overline{x})^2$, then we have the following unbiased estimators for $V(\hat{\beta}_i)$:

$$V(\hat{\beta}_1) \leftrightarrow S^2 \frac{1}{S_{xx}}; \quad V(\hat{\beta}_0) \leftrightarrow S^2 \frac{\sum x_i^2}{nS_{xx}} \tag{r9}$$

(S^2 was defined in ($r8$).) We will use as estimators for $\sigma(\hat{\beta}_i)$:

$$\sigma(\hat{\beta}_1) \leftrightarrow S\sqrt{\frac{1}{S_{xx}}}; \quad \sigma(\hat{\beta}_0) \leftrightarrow S\sqrt{\frac{\sum x_i^2}{nS_{xx}}} \tag{r10}$$

One can argue using the Central Limit Theorem that in the case of large samples, the random variables

$$\frac{\hat{\beta}_1 - \beta_1}{S\sqrt{\frac{1}{S_{xx}}}}; \qquad \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{\sum x_i^2}{nS_{xx}}}} \tag{r11}$$

may in many cases taken to be approximately $N(0,1)$ and use this to construct confidence intervals and as a basis of hypothesis tests. However, the precise conditions under which this is justified seem a bit problematical, and in addition one is often interested in the small sample situation. For this reason, it is commonly assumed that the e_i are normal. With this additional assumption, it is not diffficult to show that the random variables in ($r11$) have t distributions with $n-2$ degrees of freedom. The proof is an application of Fisher's Lemma. We give the details below.

We also note that if we just wish to test the hypothesis $H_o : \beta_1 = 0$, this can be done as follows. Under this hypothesis, the random variable

$$\frac{\hat{\beta}_1}{S\sqrt{\frac{1}{S_{xx}}}}$$

has a t distribution with $n-2$ degrees of freedom. The square of this random variable which is easily seen to be

$$\frac{SSY - SSE}{SSE/(n-2)}$$

(see ($r8$)) thus has an F distribution with 1 numerator and $n-2$ denominator degrees of freedom. Thus for a test of level α, we reject H_o if this statistic has, for the sample being tested, a value greater than $F^1_{n-2,\alpha}$ (the appropriate percentage point for the F distribution; *i.e.* $P(F^1_{n-1} > F^1_{n-2,\alpha}) = \alpha$, where F^1_{n-1} denotes a random variable whose distribution is F with 1 and $n-2$ degrees of freedom.)

We now sketch the argument that the random variables in ($r11$) have t distributions when the e_i are normal.

A somewhat longwinded calculation shows that

$$\sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \beta_0 - \beta_1 x_i)^2 - n(\overline{Y} - \beta_0)^2 - \sum (x_i - \overline{x})^2 (\hat{\beta}_1 - \beta_1)^2 \qquad (r12)$$

The variables $U_i = Y_i - \beta_0 - \beta_1 x_i$ are independent $N(0, \sigma^2)$, and if we put

$$W_1 = \sqrt{n}(\overline{Y} - \beta_0) = \frac{1}{\sqrt{n}} \sum U_i$$

$$W_2 = \sqrt{S_{xx}}(\hat{\beta}_1 - \beta_1) = \frac{1}{\sqrt{S_{xx}}} \sum (x_i - \overline{x}) U_i$$

then $(r12)$ becomes

$$\sum (Y_i - \hat{Y}_i)^2 = \sum (U_i)^2 - W_1^2 - W_2^2. \qquad (r13)$$

Since the vectors $\frac{1}{\sqrt{n}}(1, \ldots, 1)$ and $\frac{1}{\sqrt{S_{xx}}}(x_1 - \overline{x}, \ldots, x_n - \overline{x})$ are orthonormal, Fisher's Lemma implies that $\overline{Y}$, $\hat{\beta}_1$, and $\sum (Y_i - \hat{Y}_i)^2$ are independent, and that $\frac{1}{\sigma^2} \sum (Y_i - \hat{Y}_i)^2$ has a χ^2 distribution with $n - 2$ degrees of freedom.