CMI/BVR Probability notes 10

climbing a peak: frequency of decimal digits:

We are now climbing a peak: if you select a number at random from $(0, 1)$ then all digits occur with equal frequency, namely, $1/10$ — frequency exists and equals $1/10$.

More precisely, suppose you pick a number at random, say, $x \in (0, 1)$. Express

$$x = \frac{x_1}{10} + \frac{x_2}{10^2} + \frac{x_3}{10^3} + \cdots$$

where each x_i is one of the digits

$$\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

Calculate the proportion of the digit 7 in the first n places:

$$p_n(x) = \frac{\#\{1 \leq i \leq n : x_i = 7\}}{n}.$$

firstly, the sequence $\{p_n(x)\}$ *converges* and
secondly, it *converges to* $1/10$.

Obviously this is false for *lots of* x — for most rational numbers this is false (why?). In fact there are uncountably many numbers for which this is false. Fill up all the decimal digits with one or two! (How many ways are there?).

We show that the event consisting of 'all numbers x for which the statement above fails' has probability zero.

Interestingly enough we are not sure if the numbers we know, like e and π have this property. We do not even know whether each digit occurs infinitely many times in their expansions. Of course, to think in our framework, you should consider $(e-3)$ and $(\pi-3)$.

We need to climb several steps to reach this peak. Since you are not sure of decimal expansion, let that be our first step.

1. Decimal expansion:

Every number $x \in (0,1)$ can be expressed as

$$x = \frac{x_1}{10} + \frac{x_2}{10^2} + \frac{x_3}{10^3} + \cdots$$

where each x_i is one of the digits $\{0,1,2,3,4,5,6,7,8,9\}$.

This is easy. Divide the interval into 10 parts

$$(0,1/10]; \quad (1/10,2/10]; \quad \cdots, (9/10,1)$$

to be named as

$$I_0, I_1, I_2, \cdots, I_9$$

We put x_1 to be the number of the interval to which x belongs. Since x and $x_1/10$ are in the same interval of length $1/10$, we immediately have

$$|x - \frac{x_1}{10}| \leq \frac{1}{10}.$$

Now divide each of the intervals I_z into sub intervals as above (not as you like)

$$I_{z,0}; I_{z,1}, \cdots, I_{z,9}$$

Let x_2 be the digit j such that $x \in I_{x_1,j}$. Since x and $(\frac{x_1}{10} + \frac{x_2}{10^2})$ are in the same interval of length $1/100$, we sees

$$|x - (\frac{x_1}{10} + \frac{x_2}{10^2})| \leq \frac{1}{10^2}$$

Having obtained n-th level intervals and n-th digits, you proceed to the $(n+1)$-th level intervals and $(n+1)$-th digit. I leave it for you to write induction argument.

This expansion is also written as

$$x = 0.x_1x_2x_3x_4\cdots\cdots\cdots$$

Suppose a number has two different expansions:

$$0.a_1a_2a_3a_4\cdots\cdots\cdots \quad = \quad x \quad = \quad 0.b_1b_2b_3b_4\cdots\cdots\cdots$$

Since they are different, there must be a position where $a_n \neq b_n$. Let k be the first such place. Also one of them must be larger than the other, say b is larger. that is

$$a_i = b_i = \alpha_i \text{ (say)} \quad \forall i < k; \qquad a_k < b_k.$$

Thus we have

$$\begin{aligned} x &= 0.\alpha_1\alpha_2\cdots\alpha_{k-1}\ a_k\ a_{k+1}a_{k+2}\cdots\cdots\cdots \\ x &= 0.\alpha_1\alpha_2\cdots\alpha_{k-1}\ b_k\ b_{k+1}b_{k+2}\cdots\cdots\cdots \end{aligned}$$

Since all the digits are at most 9 we see

$$x \leq \sum_{i<k}\frac{\alpha_i}{10^i} + \frac{a_k}{10^k} + \sum_{i>k}\frac{9}{10^i} = \sum_{i<k}\frac{\alpha_i}{10^i} + \frac{a_k+1}{10^k} \qquad (\bullet)$$

Also each $b_i \geq 0$ gives us

$$x \geq \sum_{i<k}\frac{\alpha_i}{10^i} + \frac{b_k}{10^k} \qquad (\bullet\bullet)$$

Thus the above two inequalities $(\bullet\bullet)$ and $(\bullet)$ give

$$\sum_{i<k}\frac{\alpha_i}{10^i} + \frac{b_k}{10^k} \leq \sum_{i<k}\frac{\alpha_i}{10^i} + \frac{a_k+1}{10^k}$$

Thus $b_k \leq a_k + 1$. But since $b_k > a_k$ we conclude that

$$b_k = a_k + 1$$

If for some $i > k$ we had $b_i \neq 0$ we get using second expansion of x;

$$x \geq \sum_{i<k}\frac{\alpha_i}{10^i} + \frac{a_k+1}{10^k} + \frac{b_i}{10^i}$$

contradicting $(\bullet)$. Thus

$$b_i = 0 \qquad \forall i > k$$

Thus

$$x = 0.\alpha_1\alpha_2\cdots\alpha_{k-1}\ (a_k+1)\ 000000\cdots\cdots\cdots$$

$$x = 0.\alpha_1\alpha_2\cdots\alpha_{k-1}\ a_k\ a_{k+1}a_{k+2}\cdots\cdots\cdots$$

giving

$$\frac{a_k+1}{10^k} = \frac{a_k}{10^k} + \frac{a_{k+1}}{10^{k+1}} + \frac{a_{k+2}}{10^{k+2}} + \cdots$$

This gives

$$a_{k+1} = a_{k+2} = a_{k+3} = \cdots\cdots = 9$$

(Through out we use that all digits are between zero and nine).

Thus if a number has two expansions then they must look like this: Both agree upto a certain place; next place they differ by exactly one; the expansion with smaller digit (at the differing place) is followed by nines and the expansion with larger digit (at the differing place) is followed by zeros. In symbols; if a number has two different expansions:

$$0.a_1a_2a_3a_4\cdots\cdots\cdots \quad = \quad x \quad = \quad 0.b_1b_2b_3b_4\cdots\cdots\cdots$$

then there exists unique $k \geq 1$ such that $a_i = b_i$ for all $i < k$ and further
either $a_k = b_k + 1$ and for all $i > k$ we have $a_i = 0$; $b_i = 9$
or $b_k = a_k + 1$ and for all $i > k$ we have $b_i = 0$; $a_i = 9$.

Further if D_n is the set of numbers for which the value of k above equals n; then the sets $D_1, D_2, D_3, \cdots$ are all finite. Clearly, their union $\cup D_n$ is a countable set and these are the only numbers having two different expansions.

finally, no number has more than two expansions.

We now climb two steps which help understanding events of probability zero.

2. union of zero probability events:

Suppose $(A_n, n \geq 1)$ is a sequence of events such that $P(A_n) = 0$ for all n. Then $P(\cup A_n) = 0$

The following events are disjoint.

$$C_1 = A_1; \quad C_2 = A_2 \cap A_1^c; \quad C_3 = A_3 \cap A_1^c \cap A_2^c$$

More generally

$$C_n = A_n \cap A_1^c \cap A_2^c \cdots\cdots \cap A_{n-1}^c$$

and hence

$$P(\cup C_n) = \sum P(C_n)$$

But

$$\cup C_n = \cup A_n \quad \text{hence} \quad P(\cup A_n) = P(\cup C_n)$$

and

$$C_n \subset A_n; \;\; P(A_n) = 0 \quad \text{hence} \quad P(C_n) = 0$$

Thus

$$P(\cup A_n) = 0$$

There is a general fact hidden in the above argument. Consider any sequence of events (A_n); then above argument shows that

$$P(\cup A_n) = P(\cup C_n) = \sum P(C_n) \leq \sum P(A_n)$$

Thus for any sequence of events; disjoint or not;

$$P(\cup A_n) \leq \sum P(A_n).$$

3. Borel-cantelli Lemma:

Suppose $(B_n : n \geq 1)$ is a sequence of events with $\sum P(B_n) < \infty$. let B be the set of all points which belong to infinitely many of the sets of the sequence. Then $P(B) = 0$.

Any point which is in infinitely many of the events must be in $\underset{n \geq k}{\cup} B_n$; whatever be k. Thus

$$B \subset \bigcup_{n=k}^{\infty} B_n \quad \text{for all} \quad k \geq 1$$

Thus from the observation made at the end of the previous step, we see

$$P(B) \leq P(\bigcup_{n=k}^{\infty} B_n) \leq \sum_{n=k}^{\infty} P(B_n)$$

The right side is the tail sum of a convergent series and hence becomes smaller and smaller as k becomes larger and larger. But since above inequality holds for all k, we conclude

$$P(B) = 0$$

We shall now climb one step which is an elementary calculation of expectation.

4. expectation of a sum:

Suppose $U_1, U_2, \cdots U_n$ are independent mean zero random variables and $|U_i| \leq 1$ for each i. Then

$$E\left[\left(\sum U_i\right)^4\right] \leq 7n^2$$

Proof is simple algebra.

$$(\sum U_i)^4 = (\sum U_i)(\sum U_i)(\sum U_i)(\sum U_i)$$

is sum of several products which can be grouped as follows (below different indices are distinct):

$$(1) \quad \sum U_i^4; \quad (2) \quad \sum U_i^2 U_j^2$$

and

$$(3) \quad \sum U_i U_j^3; \quad (4) \quad \sum U_i U_j U_k^2; \quad (5) \quad \sum U_i U_j U_k U_l$$

By independence

$$E(U_i U_j^3) = E(U_i)E(U_j^3) = 0$$

Similarly expectations of the other terms appearing in (4) and (5) are also zero. Thus no matter how many terms are there in these sums (3,4,5) their expectation is zero.

Use now $|U_i| \leq 1$ for each i to conclude $E(U_i)^4 \leq 1$ and there are n summands in (1).

Similarly $E(U_i^2 U_j^2) \leq 1$ and there are at most $6n^2$ summands in (2). This is because

out of the four factors in $(\sum U_i)^4$ take U_i from two and U_j from the remaining two — can do in six ways;

also i and j can be selected in at most n^2 ways (actually less, since the indices are distinct, you see $\binom{n}{2}$ ways, but why bother!).

Thus expectations of (1) and (2) above add to at most

$$n + 6n^2 \leq 7n^2.$$

Now we climb a step which explains handling the decimal digits of the selected point.

5. decimal digits of point selected at random:

Let X be a point selected at random from $(0,1)$. Let X_i be the i-th decimal digit of X. Thus for example; X_1 equals $0,1,2,\cdots,9$ according as X is in the intervals $I_0, I_1, \cdots, I_9$. We are using notation of step one. Thus the event $(X_1 = k)$ is same as $(X \in I_k)$. Since I_k has length $1/10$ we conclude that X_1 takes values $0,1,2\cdots,9$ with probability $1/10$ each.

X_2 also takes the same values. We see that

$$(X_2 = k) = \bigcup_{j=0}^{9} (X \in I_{jk})$$

The intervals I_{jk} are disjoint and each has length $1/100$ and thus

$$P(X_2 = k) = \frac{1}{10} \qquad k = 0,1,2,\cdots,9.$$

In general you see that the event $(X_n = k)$ is union of 10^{n-1} events each of probability $1/10^n$ and hence

$$P(X_n = k) = \frac{1}{10} \qquad k = 0,1,2,\cdots,9.$$

We now argue that for every n; the random variables $X_1, \cdots, X_n$ are independent. Indeed if you take any digits $k_1, \cdots, k_n$ then the event

$$(X_1 = k_1, X_2 = k_2, \cdots, X_n = k_n)$$

consists of the interval

$$I_{k_1 k_2 \cdots k_n}$$

at the n-th level and has length $1/10^n$. Thus

$$P(X_1 = k_1, X_2 = k_2, \cdots, X_n = k_n) = \frac{1}{10^n} = \prod_1^n P(X_i = k_i)$$

Thus the decimal digits of a point selected at random are independent and each has the same distribution described above.

We shall now slowly enter problem of interest. We climb a step which helps understand occurrence of the digit 7.

6. Digit 7 occurrence:

Let X and all X_i be as above. define Z_i for $i \geq 1$ as follows. $Z_i = 1$ if the i-th decimal digit of X is 7; otherwise $Z_i = 0$. Thus Z_i indicates whether the i-th digit is 7 or not.

In terms of the X_i we can express:
$Z_i = 1$ if $X_i = 7$; and $Z_i = 0$ if $X_i \neq 7$.

Since Z_i depends only on X_i we see that for any n, the random variables $Z_1, Z_2, \cdots, Z_n$ are independent. Remember $X_1, X_2, \cdots, X_n$ are independent.

Also each Z_i has the same distribution: one with probability $1/10$ and zero with probability $9/10$.

Let us note that

$$E(Z_i) = \frac{1}{10}$$

We shall now get closer to our quantity of interest. Observe that

$$Z_1 + Z_2 + \cdots + Z_n$$

counts precisely the number of occurrences of the digit 7 in the first n decimal places of X. Hence $(Z_1 + Z_2 + \cdots + Z_n)/n$ counts the proportion of times digit 7 appears in the first n decimal places of X. The next step tells us how close this is to $1/10$.

7. Proportion of digit 7 in first n places:

For each n;

$$E\left\{\frac{Z_1 + \cdots + Z_n}{n} - \frac{1}{10}\right\}^4 \leq \frac{7}{n^2}$$

Also for each $\epsilon > 0$

$$\sum_{n=1}^{\infty} P\left\{\left|\frac{Z_1 + \cdots + Z_n}{n} - \frac{1}{10}\right| > \epsilon\right\} < \infty.$$

To see these, first recall Chebyshev inequality: for a non-negative random variable W and $\epsilon > 0$

$$P(W > \epsilon) = P(W^4 > \epsilon^4) \leq E(W^4)/\epsilon^4$$

Taking

$$W = \left| \frac{1}{n} \sum Z_i - \frac{1}{10} \right|$$

and noting that the series $\sum (1/n^2)$ is convergent you see that the first claim above implies the second. Remember $\epsilon > 0$ is fixed. we need to prove the first claim now.

Since Z_i are independent, we conclude $\{Z_i - \frac{1}{10} : 1 \leq i \leq n\}$ are independent, they have mean zero, they take values

$$0 - \frac{1}{10}; \qquad 1 - \frac{1}{10}$$

and hence are at most one in magnitude. Note that the quantity in brackets is just

$$\frac{(Z_1 - 1/10) + (Z_2 - 1/10) + \cdots + (Z_n - 1/10)}{n}$$

Thus step 4 gives the required expectation is at most

$$\frac{7n^2}{n^4} = \frac{7}{n^2}.$$

Now that we have got an idea of how far the proportion is from 1/10, we take the final step to reach the peak.

8. Proportion of the digit 7:

We now show:

$$P\left\{ \frac{Z_1 + \cdots + Z_n}{n} \not\to \frac{1}{10} \right\} = 0.$$

Let S_1 be the event

$$S_1 \quad = \quad \left\{ \left| \frac{Z_1 + \cdots + Z_n}{n} - \frac{1}{10} \right| > 1 \quad \text{for infinitely many} \quad n \right\}$$

If you take $\epsilon = 1$ in the previous step and use step 3 (Borel-Cantelli) you get

$$P(S_1) = 0$$

Let S_2 be the event

$$S_2 \quad = \quad \left\{ \left| \frac{Z_1 + \cdots + Z_n}{n} - \frac{1}{10} \right| > \frac{1}{2} \quad \text{for infinitely many} \quad n \right\}$$

If you take $\epsilon = 1/2$ in the previous step and use step 3 (Borel-Cantelli) you get

$$P(S_2) = 0$$

More generally, Let S_k be the event

$$S_k \quad = \quad \left\{ \left| \frac{Z_1 + \cdots + Z_n}{n} - \frac{1}{10} \right| > \frac{1}{k} \quad \text{for infinitely many} \quad n \right\}$$

If you take $\epsilon = 1/k$ in the previous step and use step 3 (Borel-Cantelli) you get

$$P(S_k) = 0$$

Now use step 2 to get

$$P(\bigcup_1^\infty S_k) = 0$$

Let

$$S = (\bigcup_1^\infty S_k)^c$$

so that

$$P(S) = 1$$

We now argue that for all points in the event S;

$(Z_1 + \cdots + Z_n)/n$ converges and it converges to $1/10$.

If you take any point $\omega \in S$ then it is not in S_k whatever be k. Thus whatever k is given to us, we conclude:

$$\left| \frac{Z_1(\omega) + \cdots + Z_n(\omega)}{n} - \frac{1}{10} \right| \quad \leq \quad \frac{1}{k} \quad \text{for all } n \text{ after some stage}$$

In other words if $\omega \in S$ then

$$\frac{Z_1(\omega) + \cdots + Z_n(\omega)}{n} \to \frac{1}{10}$$

You would quickly realize that the digit 7 played no role, any digit would lead to the same conclusion. Thus we have proved:

For almost all numbers in $(0, 1)$ *each digit occurs with frequency 1/10 in its decimal expansion.*

The above sentence mean the following: if you select a number at random from $(0,1)$ then with probability one, frequency of each digit in decimal expansion equals $1/10$. The phrase 'almost all' refers to the fact that the event for which the stated property fails is 'unimportant' 'small' or or be precise, 'has probability zero'.

This is easy to deduce from the above. Let X be a number selected at random from $(0,1)$. Put for $0 \leq i \leq 9$;

$$A_i = \{\omega : \text{In the decimal expansion of } X(\omega) \text{ digit } i \text{ has frequency } 1/10\ \}$$

The above argument tells $P(A_i) = 1$ for each i. Thus if $A = \cap A_i$ then $P(A) = 1$ and for $\omega \in A$, the decimal expansion of $X(\omega)$ has *each* digit occuring with frequency $1/10$.

other patterns:

Let us fix a pattern, that is, fix an integer $k \geq 1$ and an ordered sequence $s = a_1 a_2 \cdots a_k$ of decimal digits. For a number $x \in (0,1)$ with decimal expansion

$$x = 0.x_1 x_2 x_3 x_4 \cdots\cdots$$

we say that the pattern s occurs at place m, if

$$(x_{m-k+1} x_{m-k+2} \cdots x_m) = a_1 a_2 \cdots a_k \quad i.e. \quad = s$$

By convention, for $m < k$, we take it that the pattern does not occur at place m. Let

$$p_n(x) = \#\{m \leq n : \text{pattern } s \text{ occurs at place } m\}/n$$

proportion of occurrence of the pattern s till place n. Then a little change in the earlier argument (careful or even careless estimate of $E[(\sum U_i)^4]$) shows that almost surely $p_n(X)$ converges to $1/10^k$. remember there are 10^k patterns of length k. In other words, denoting $|s|$ for length of s; and denoting

$$B_s = \{\omega : p_n(X(\omega)) \not\to 1/10^{|s|}\}$$

we have

$$P(B_s) = 0.$$

Since for each k there are finitely many patterns of length k we see that there are countable many patterns. Thus

$$P(\bigcup_s B_s) = 0$$

Here the union is over all patterns of all possible (finite) lengths. If you fix any point ω in the complement of the set above you see that for the number $X(\omega)$ *each* pattern occurs with right frequency. Thus we have the following.

For almost all numbers in $(0,1)$*, the following is true: what ever finite decimal pattern s you take it occurs with frequency* $1/10^{|s|}$.

Other expansions:

Let $r > 1$ be an integer. We can (you Should) expand $x \in (0,1)$ with base r.

$$x = \frac{x_1}{r} + \frac{x_2}{r^2} + \frac{x_3}{r^3} + \frac{x_4}{r^4} + \cdots\cdots$$

where for each i; $x_i \in \{0,1,2,3,\cdots,r-1\}$.

You can (Should) also show that only countably many numbers have two expansions and no number has more than two expansions. Exactly the same proof above with *no change* shows the following.

For almost all numbers in $(0,1)$*, in its r-expansion; each of the digits* $\{0,1,2,\cdots,(r-1)\}$ *occurs with frequency* $1/r$.

(Earn the right to proceed further)

In fact more is true. Let us now define a r-pattern, that is, fix an integer $k \geq 1$ and an ordered sequence $s = a_1a_2\cdots a_k$ of digits $\{0,1,2,\cdots(r-1)\}$. For a number $x \in (0,1)$ with r-expansion

$$x = 0.x_1x_2x_3x_4\cdots\cdots$$

we say that the pattern s occurs at place m, if

$$(x_{m-k+1}x_{m-k+2}\cdots x_m) = a_1a_2\cdots a_k \quad i.e. \quad = s$$

By convention, for $m < k$, we take it that the pattern does not occur at place m. Let

$$p_n(x) = \#\{m \leq n : \text{pattern } s \text{ occurs at place } m\}/n$$

proportion of occurrence of the pattern s till place n. Then we can easily show that almost surely $p_n(X)$ converges to $1/r^k$. remember there are r^k patterns of length k. In other words, denoting $|s|$ for length of s; and denoting

$$B_s = \{\omega : p_n(X(\omega)) \not\to 1/r^{|s|}\}$$

we have

$$P(B_s) = 0.$$

Since for each k there are finitely many patterns of length k we see that there are countable many patterns. Thus

$$P(\bigcup_s B_s) = 0$$

If you fix any point ω in the complement of the set above you see that for the number $X(\omega)$ *each* pattern occurs with right frequency. Thus we have the following.

For almost all numbers in $(0,1)$*, the following is true: what ever finite* r*-pattern* s *you take it occurs with frequency* $1/r^{|s|}$.

absolutely normal numbers:

Let us say that a number x is absolutely normal if the following holds: if you take an integer $r > 1$ and if you take a r-pattern s; then s occurs with frequency $1/r^{|s|}$ in the r-expansion of x. In other words each pattern in each base should occur with the right frequency in expansion to that base. Using countable union of events of probability zero has again probability zero, we have the following.

Almost all numbers in $(0,1)$ *are absolutely normal.*

That is, if you select a number at random from $(0,1)$ then it is absolutely normal. This profound theorem is essentially due to Borel.

The most interesting point is that we have very very few concrete examples of such numbers. We do not know whether the numbers familiar to us; like e or π *are absolutely normal. Are they normal at least in decimal expansion? Do not know. Forget about right frequency; would zero occur infinitely many times in their decimal expansion. I do not know; probably no one knows.*

Life is like that!

CLT again:

In earlier discussion our experiment was selecting a number at random from $(0,1)$, a continuous random variable. We have explained an aspect of such a random number; but through out the analysis we used discrete random variables. In other words discrete random variables helped understand a continuous random variable.

We shall now express some probabilities involving discrete random variables with the help of continuous random variable. In other words a continuous random variable helps understanding discrete random variables. It is CLT. We have already seen a special case of this due to DeMoivre and Laplace — coin tossing.

CLT:

Suppose we have a sequence of discrete random variables: $X_1, X_2, \cdots$.

Assume that they all have same distribution.

Assume that $E(X_1) = \mu$ and variance$(X_1) = \sigma^2$ with $0 < \sigma^2 < \infty$.

Assume that for each n; the random variables $X_1, X_2, \cdots, X_n$ are independent.

Then for any two numbers $a < b$

$$P\{a < \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} < b\} \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

as $n \to \infty$.

Note that the left side is probability of a discrete random variable. The limit is standard normal variable probability.

In other words sum of a large number of random variables, when normalized, looks like a standard normal variable. If you denote

$$S_n = X_1 + X_2 + \cdots + x_n$$

then

$$\frac{S_n - n\mu}{\sqrt{n}\sigma}$$

is nothing but standardized sum. Recall standardisation means making mean zero and variance one; achieved by subtracting its mean and dividing by its standard deviation.

You should note that nothing is assumed about the distribution (except variance nonzero and finite).

For example if each X_i is Bernoulli(p) then

$$S_n \sim B(n, p)$$

and the above theorem reduces to the DeMoivre-Laplace theorem.

If you take $X_i \sim P(1)$ then $S_n \sim P(n)$ and you have a similar theorem. Of course, you can take the common distribution to be anything reasonable.

There is another way of looking at this theorem. Instead of saying that we are looking at the standardized S_n we can look at it as follows:

$$\frac{X_1 - \mu}{\sqrt{n}\sigma} + \frac{X_2 - \mu}{\sqrt{n}\sigma} + \frac{X_3 - \mu}{\sqrt{n}\sigma} + \cdots + \frac{X_n - \mu}{\sqrt{n}\sigma}$$

it is sum of n things; each thing has $\sqrt{n}$ in the denominator and so very small (as n gets larger). Thus the expression can be thought of as 'sum of a large number of variables, each of which is very small'. Looked at this way, it says that

a large number of very very small random variables (errors?) is approximately normal.

When you study linear models or regression techniques and assume that errors of measurements are normal, the above provides a theoretical justification. Usually measurement error is due to several factors, each factor contributing a little towards the overall error. We shall not enter into the details.

Actually the theorem is true without the word 'discrete'. Thus you can take any random variables. We return to this after we define independence in the continuous case.

density and DF:

Suppose X is standard normal. that is, has density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad\qquad -\infty < x < \infty$$

Remember, this means that for any interval (a, b)

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Let us consider the following function:

$$g(x) = f(x) \quad x \neq 0; \qquad g(0) = 10^{10}$$

Properties of Riemann integral should tell you that for any $a < b$

$$\int_a^b g(x)dx = \int_a^b f(x)dx$$

In other words we are perfectly right if we said X has density g.

($\bullet$) Thus density is not unique. You can change at all integer points. You can change at many points; but the point is you can only change on a set which is in some sense small, some sense unimportant, in some sense irrelevant. We shall not expand on this theme.

Whenever a 'beautiful' density exists we use that rather than an 'ugly' density. For example g above is ugly and f is beautiful.

If you know the density you can calculate DF. Simply define

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

Remember, by definition of DF, $F(a) = P(X \leq a)$.

Conversely, if you have a random variable; if you know it has density; but you know only DF; can you recover density from DF?

($\bullet\bullet$) Yes. Let F be the distribution function of X. Define for $x \in R$; $f(x)$ as the derivative of F at x, namely, $F'(x)$ when it exists. If at a point x, F is not differentiable, put $f(x) = 0$ This function will be density. In other words

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

holds for f as defined above.

However you should not blindly use this without verifying that X indeed has density. If you did not know this, you can still define f as above, but then verify directly that this works as density for X.

Interesting case is the Cantor distribution function F. Suppose X is a random variable with this DF. Note that complement of Cantor set is a countable union of intervals and on each of those intervals F is a constant (depending on the interval). In particular derivative of F exists on complement of Cantor set and equals zero! It can be shown, we shall not do, that there is NO density in this case. that is, a random variable X having this DF, has no density.

If X has density f then definition of density already tells you that for any interval (a, b) we have

$$P(a < X < b) = \int_a^b f(x)dx$$

Is this true for more general sets, not just intervals? that is

$$P(X \in A) = \int_A f(x)dx$$

$(\bullet \bullet \bullet)$ Yes. But you need to be careful because Riemann integral is defined 'over' intervals, finite unions etc. Integral over general A is not defined. Lebesgue integral is defined and the above equality does indeed hold. But we shall not get into this, we do not need in our course.

independence:

For two random variables X and Y, in the discrete case we said that they are independent if for any two numbers a and b

$$P(X = a, Y = b) = P(X = a)P(Y = b). \quad (\clubsuit)$$

If X and Y have densities then such a definition does not make sense because both sides are zero and the above equality always holds.

In the discrete case, the above definition is equivalent to the following: For any two intervals I and J,

$$P(X \in I; Y \in J) = P(X \in I)P(Y \in J). \qquad (\spadesuit)$$

Indeed, if $(\spadesuit)$ is granted and if you fix a and b to verify $(\clubsuit)$ proceed as follows. Apply hypothesis for the intervals $I = (a - \frac{1}{n}, a + \frac{1}{n})$ and $J = (b - \frac{1}{n}, b + \frac{1}{n})$ and take limit over n on both sides of the equality. remember, if events A_n decrease to A then probabilities decrease too.

Conversely, if (♣) is granted and if you fix intervals I, J and want to verify (♠) for intervals I and J proceed as follows. Apply hypothesis for each $a \in I$ and $b \in J$ such that $P(X = a) > 0$ and $P(Y = b) > 0$. add the countably many equalities. Remember, X being discrete $P(X \in I)$ is same as $\sum P(X = a)$ where sum is over $a \in I$ with $P(X = a) > 0$.

Two random variables X, Y defined on a space are said to be independent if for any two intervals I and J

$$P(X \in I; Y \in J) = P(X \in I)P(Y \in J).$$

It is a matter of technical detail whether you demand the above to hold for all open intervals I, J; or for all closed intervals; or for all intervals of the form leftopen-rightclosed; or for all intervals with rational end points; or for only intervals of the form $I = (-\infty, a]; J = (-\infty, b]$. Any one of these demands meets all the others as well!

integration; several variables:

You are doing higher level topics in calculus and I need down-to-earth integration. So I shall review integration little bit. But please keep in mind what I define is neither the usual definition nor the correct one. Then why do I follow this road? This is correct and enough as far as functions we come across in this course are concerned. Subtle details, part of calculus, are not our concern. You should be under no illusion that we are developing integration.

suppose $f : R^2 \to R$. we define $\int\int f(x,y)dxdy$ as follows. First, for each fixed y, consider the function $x \mapsto f(x,y)$ from R to R and integrate it. You get a number; but this number depends on which y you fixed and so we denote this number by $\varphi(y)$. Now integrate thus function: $\int \varphi(y)dy$. You get a number. This is the meaning of the above integral. In other words

$$\int\int f(x,y)dxdy = \int \left[\int f(x,y)dx\right] dy.$$

Of course you can similarly define by interchanging the order

$$\int\int f(x,y)dydx = \int \left[\int f(x,y)dy\right] dx.$$

For the functions we come across in our course, both give same answer and this is taken as

$$\int_{R^2} f(x,y)d(x,y)$$

called 'double integral' or integral of f over R^2. the earlier two are called as 'repeated integrals', simply because we are calculating integrals one after another there.

Just like the definition of Riemann integral for functions of one variable, we can define Riemann integral for functions of two variables and that is called double integral. In general,
such a double integral need not equal the earlier repeated integrals;
even those two repeated integrals may lead to different values.

However, considering the fact that we wish to do probability and not calculus; considering the fact that calculus is only a tool for us; considering the fact that functions that we discuss do not lead to the pathologies mentioned; we have taken this route. But what you should appreciate is that we explained matters only using known concept of integration of functions of one variable.

Example:

$f(x, y)$ is indicator of $D = \{(x, y) : x^2 + y^2 \leq 1\}$ That is, takes values one or zero according as the point is in D or not. If you fix y and integrate w.r.t x you get the following: In case $|y| > 1$ then for every x, you see $f(x, y) = 0$ and hence integral is zero. In case $|y| \leq 1$, then you get $2\sqrt{1 - y^2}$. Thus

$$\varphi(y) = 2\sqrt{1 - y^2} \quad |y| \leq 1; \qquad \varphi(y) = 0 \text{ for } \quad |y| > 1$$

And

$$\int \varphi(y) dy = \pi$$

Thus if

$$g(x, y) = \frac{1}{\pi} f,$$

then

$$\int g dx dy = 1.$$

Example:

$f(x, y)$ is 2 when $0 < y < x < 1$ and zero otherwise. Thus this indicator of the lower half of the unit square. You see $\varphi(y) = 2(1 - y)$ if $0 < y < 1$ and $\varphi(y) = 0$ otherwise. Finally $\int \varphi = 1$.

Example:

$$f(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}$$

Then

$$\int\int f dx dy = 1$$

You can similarly define integration of function of any number of variables, integrate variables one by one. For example if you have a function of 20 variables $f(x_1, x_2, \cdots, x_{20})$ then for each fixed 19-tuple $(x_1, \cdots, x_{19})$ this is a function of only one variable, namely function of x_{20} only. Integrate it to get $f_{19}(x_1, \cdots, x_{19})$. Now for each fixed 18-tuple $(x_1, \cdots, x_{18})$ this is a function of one variable, namely function of x_{19}; integrate w.r.t. x_{19} to get $f_{18}(x_1, \cdots, x_{18})$. continue till you get $f_1(x_1)$ and integrate it to get $\int f_1(x_1)dx_1$, a number. This is the integral of f over R^{20} or integral of f w.r.t. all the variables.

Example: Consider R^{20}.

$$f(x) = \frac{1}{(2\pi)^{10}} e^{-x^t x/2} = \frac{1}{(2\pi)^{10}} e^{-\sum x_i^2/2}; \qquad x \in R^{20}$$

You see

$$\int_{R^{20}} f(x) dx = 1$$

Of course we have used dx to mean $dx_1 dx_2 \cdots dx_{20}$.

We have been able to calculate integral of even a function of twenty variables.

There is only one subtle point. Suppose that you have $f(x, y)$. After integrating w.r.t. x you get $\varphi(y)$. We said integrate this w.r.t. y. *question:* how do you know this is integrable? To make sense of the prescription, you need to ensure that this is integrable, otherwise the prescription can not be implemented, the algorithm stops in the middle.

However, our functions will not lead to this situation. So we do not enter the theoretical realm. We are happy we can do what ever we want to do. It is all that matters.

Joint densities:

Suppose X and Y are two random variables defined on a probability space.

We say that a function $f : R^2 \to [0, \infty)$ is joint density of (X, Y) if the following holds: for any $a < b$ and $c < d$

$$P(a < X < b; c < Y < d) = \int_{(a,b)\times(c,d)} f(x, y)dxdy$$

In other words, the chances that the pair (X, Y) is in a rectangle equals integral of f over the rectangle. This is exactly like one random variable: probability that a random variable X is in an interval equals integral over that interval of the density of X. Remember

$$\int_{(a,b)\times(c,d)} f(x, y)dxdy = \int_c^d \left[\int_a^b f(x, y)dx\right] dy$$

$$= \int_a^b \left[\int_c^d f(x, y)dy\right] dx$$

A point is to be carefully noted here. Earlier when wrote repeated integrals, they were on all of R, so limits of integration did not show up, they are understood as $-\infty$ to $+\infty$. But notice how limits are showing up here and when you change the order, the limits accordingly change. Nothing new. You may recall double sums: if i is from 1 to 10 and j is from 4 to 14 and when you interchange order of summation what happens?

An interesting question: Suppose X and Y are defined on one probability space and they have densities. Then does (X, Y) have joint density? Not necessarily.

Suppose $X \sim Unif(0, 1)$ and $Y = X$. Thus the pair (X, Y) always takes values on the diagonal, after all Y is same as X. That is if

$$D = \{(a, a) : 0 < a < 1\}$$

then

$$P\{(X, Y) \in D\} = 1.$$

One can argue that there is no joint density for the pair (X, Y) even though each of X and Y have densities. We shall not spend time on this, it is not difficult.

In one special case the above question has affirmative answer. Suppose X and Y have densities $f(x)$ and $g(y)$. If they are independent then (X,Y) has joint density and it is given by

$$h(x,y) = f(x)g(y).$$

Indeed

$$\int_c^d \left[\int_a^b h(x,y)dx\right] dy = \int_c^d \left[\int_a^b f(x)g(y)dx\right] dy$$

$$= \int_c^d P(a < X < b) \;\; g(y)dy = P(a < X < b)P(c < Y < d)$$

$$= P(a < X < b; c < Y < d)$$

where we used independence in the last step. This shows that h is a joint density for (X,Y).

If (X,Y) have joint density, does X have density? Yes. Let $f(x,y)$ be the joint density of (X,Y). Put

$$f_X(x) = \int f(x,y)dy$$

Then we claim that f_X is density of X. Indeed

$$P(a < X < b) = P(a < X < b; -\infty < Y < \infty)$$

$$= \int_a^b \left[\int_{-\infty}^{\infty} f(x,y)dy\right] dx = \int_a^b f_X(x)dx$$

This f_X is called marginal density of X. Keep in mind marginal density is nothing but density. the adjective 'marginal' is just to draw your attention to the fact there are others floating around and you calculated this density using the joint density.

Similarly you can see that Y has the following density

$$f_Y(y) = \int f(x,y)dx$$

I hope you have spotted the analogy with discrete set up. For two discrete random variables (X,Y) their joint distribution is presented as a bivariate table with entries

$$p_{ij} = P(X = x_i, Y = y_j)$$

The marginal density of X is defined by

$$P(X = x_i) \quad = \quad \sum_j p_{ij} \quad = \quad p_{i\bullet}$$

The change in going from discrete to continuous case:

(i, j) is replaced by (x, y)

p_{ij} is replaced by $f(x, y)$

$\sum\limits_j p_{ij}$ is replaced by $\int f(x, y)dx$

$p_{i\bullet}$ is replaced by $f_X(x)$.

You can push matters further. Remember we defined conditional distribution of X given $Y = y_b$ by using the numbers $p_{ib}/p_{\bullet b}$.

Suppose you fix an b such that $f_Y(b) > 0$. Then you can define the conditional density of X given $Y = b$ as follows

$$f_{X|(Y=b)}(x) = \frac{f(x, b)}{f_Y(b)}$$

This is well defined because the denominator is assumed to be non-zero. This integrates (x is the variable, b is fixed) to one and can legitimately be called conditional density of X given $Y = b$.

Just keep in mind that in the above expression b is fixed; it is regarded as a function of x and it is a density function as a function of x.

We can continue with conditional expectation and so on. These are not just for the sake of generalizing to the continuous case; these are useful concepts.

However, we shall proceed in another direction.