Frequency distribution activity

Goal: Turn the following cipher text into plaintext, using frequency distribution. It is known that the ciphertext and plaintext are both written using the English language.

yfzh zh mj kpktlzhk zj etkmizjq lzwfkth dhzjq vtkcdkjlg bzhytzedyznj. zy zh kmhzkhy yn dhk yfzh skyfnb zv gnd zjyktlkwy m skhhmqk yfmy zh sntk yfmj njk fdjbtkb lfmtmlykth zj akjqyf, eklmdhk zj yfmy lmhk, yfk hmswak hzuk zh amtqk kjndqf hn yfmy yfk lfmtmlykt bzhytzedyznj zh sntk azikag yn lanhkag tkhkseak yfmy nv m ygwzlma skhhmqk xtzyykj zj kjqazhf. xk fmok smbk gndt ymhi kmhzkt zj yfzh kpmswak eg hkwmtmyzjq yfk xntbh zj yfk skhhmqk.

The message is long enough to warrant attempting decryption using frequency distribution. The first step is to make a list of the frequency distribution of the letters in the ciphertext. It is as follows:

a - 11

b - 7

c - 1

d - 10

e - 6

f - 21

g - 7

h - 35

i - 3

j - 21

k - 52

l - 12

m - 29

n - 15

o - 1

p - 2

q - 10

r - 0

s - 10

t - 21

u - 1

v - 3

w - 6

x - 3

y - 35

z - 35

The following is an approximation of the distribution of letters in English, given a random writing sample of 1000 characters:

A - 73

B - 9

C - 30

D - 44

E - 130

F - 28

G - 16

H - 35

I - 74

J - 2

K - 3

L - 35

M - 25

N - 78

O - 74

P - 27

Q - 3

R - 77

S - 63

T - 93

U - 27

V - 13

W - 16

X - 5

Y - 19

Z - 1

In the ciphertext above, the letter k appears most frequently, with 52 instances, and the letters h, y and z are next, with 35 instances each. This implies that the cipher “k” probably translates to the plaintext “e,” since “e” is the most commonly-found letter in English, and the cipher letters “h,” “y” and “z” probably translate into three of the letters “a,” “i,” “n,” “o,” “r,” “s” or “t,” since these are the next-most-common letters in English.

Changing the ciphertext “k” to the plaintext “E,” we have the following:

yfzh zh mj EpEtlzhE zj etEmizjq lzwfEth dhzjq vtEcdEjlg bzhytzedyznj. zy zh EmhzEhy yn dhE yfzh sEyfnb zv gnd zjyEtlEwy m sEhhmqE yfmy zh sntE yfmj njE fdjbtEb lfmtmlyEth zj aEjqyf, eElmdhE zj yfmy lmhE, yfE hmswaE hzuE zh amtqE Ejndqf hn yfmy yfE lfmtmlyEt bzhytzedyznj zh sntE aziEag yn lanhEag tEhEseaE yfmy nv m ygwzlma sEhhmqE xtzyyEj zj Ejqazhf. xE fmoE smbE gndt ymhi EmhzEt zj yfzh EpmswaE eg hEwmtmyzjq yfE xntbh zj yfE sEhhmqE.

A study of short words (two or three letters) comes in handy here. We notice a few patterns, especially in reference to the most common letters seen in this ciphertext. Since "z" and "h" appear so frequently, and we notice there are five instances of the two-letter word "zh," a good guess is that "zh" could be "IS," "IN," "AT," "AN" or "OR." Also, there are six times where "zj" appears, giving more strength to this argument. Let's try the cipher "z" corresponding to the plaintext "I," with the cipher "h" corresponding to the plaintext "S" and the cipher "j" corresponding to the plaintext "N." Then we have:

yfIS IS mN EpEtlISE IN etEmiINq lIwfEtS dSINq vtEcdENlg bISytIedyInN. Iy IS EmSIESy yn dSE yfIS sEyfnb Iv gnd INyEtlEwy m sESSmqE yfmy IS sntE yfmN nNE fdNbtEb lfmtmlyEtS IN aENqyf, eElmdSE IN yfmy lmSE, yfE SmswaE SIuE IS amtqE ENndqf Sn yfmy yfE lfmtmlyEt bISytIedyInN IS sntE aIiEag yn lanSEag tESEseaE yfmy nv m ygwIlma sESSmqE xtIyyEN IN ENqaISf. xE fmoE smbE gndt ymSi EmSIEt IN yfIS EpmswaE eg SEwmtmyINq yfE xntbS IN yfE sESSmqE.

Looking at the first two words, "yfIS IS," one might guess that this means "this is," especially with the cipher "y" appearing 35 times. Guessing that provides us with:

THIS IS mN EpEtlISE IN etEmiINq lIwHEtS dSINq vtEcdENlg bISTtIedTInN. IT IS EmSIEST Tn dSE THIS sETHnb Iv gnd INTEtlEwT m sESSmqE THmT IS sntE THmN nNE HdNbtEb lHmtmlTEtS IN aENqTH, eElmdSE IN THmT lmSE, THE SmswaE SIuE IS amtqE ENndqH Sn THmT THE lHmtmlTEt bISTtIedTInN IS sntE aIiEag Tn lanSEag tESEseaE THmT nv m TgwIlma sESSmqE xtITTEN IN ENqaISH. xE HmoE smbE gndt TmSi EmSIEt IN THIS EpmswaE eg SEwmtmTINq THE xntbS IN THE sESSmqE.

Looking at the second line, we find a one-letter word "m." Since the plaintext "I" is already used, this must mean the cipher "m" corresponds to the plaintext "A." This gives:

THIS IS AN EpEtlISE IN etEAiINq lIwHEtS dSINq vtEcdENlg bISTtIedTInN. IT IS EASIEST Tn dSE THIS sETHnb Iv gnd INTEtlEwT A sESSAqE THAT IS sntE THAN nNE HdNbtEb lHAtAlTEtS IN aENqTH, eElAdSE IN THAT lASE, THE SAswaE SIuE IS aAtqE ENndqH Sn THAT THE lHAtAlTEt bISTtIedTInN IS sntE aIiEag Tn lanSEag

tESEseaE THAT nv A TgwllAa sESSAqE xtITTEN IN ENqaISH. xE HAoE sAbE gndt TASi EASIEt IN THIS EpAswaE eg SEwAtATINq THE xntbS IN THE sESSAqE.

In the second line, there is a two-letter word, “Tn,” which implies that the cipher “n” is the plaintext “O.” Then, in the fourth line, the word “ENndqH” implies “ENOUGH,” so that the cipher “d” is the plaintext “U,” and the cipher “q” is the plaintext “G.” When we put in those three substitutions, we get:

THIS IS AN EpEtlISE IN etEAiING llwHEtS USING vtEcUENlg bISTtIeUTION. IT IS EASIEST TO USE THIS sETHOb lv gOU INTEtlEwT A sESSAGE THAT IS sOtE THAN ONE HUNbtEb lHAtAlTEtS IN aENGTH, eElAUSE IN THAT lASE, THE SAswaE SluE IS aAtGE ENOUGH SO THAT THE lHAtAlTEt bISTtIeUTION IS sOtE aliEag TO laOSEag tESEseaE THAT Ov A TgwllAa sESSAGE xtITTEN IN ENGaISH. xE HAoE sAbE gOUt TASi EASIEt IN THIS EpAswaE eg SEwAtATING THE xOtbS IN THE sESSAGE.

It’s getting much easier now, because we can see obvious words formed. For instance, in the third line, “HUNbtEb” implies “HUNDRED,” in the fifth line, “ENGaISH” implies “ENGLISH,” and in the last line, “sESSAGE” implies “MESSAGE.” When we make those substitutions, we see:

THIS IS AN EpERlISE IN eREAiING llwHERS USING vREcUENlg DISTRIeUTION. IT IS EASIEST TO USE THIS METHOD lv gOU INTERlEwT A MESSAGE THAT IS MORE THAN ONE HUNDRED lHARAlTERS IN LENGTH, eElAUSE IN THAT lASE, THE SAMwLE SluE IS LARGE ENOUGH SO THAT THE lHARAlTER DISTRIeUTION IS MORE LliELg TO lLOSELg RESEMeLE THAT Ov A TgwllAL MESSAGE xRITTEN IN ENGLISH. xE HAoE MADE gOUR TASi EASIER IN THIS EpAMwLE eg SEwARATING THE xORDS IN THE MESSAGE.

Although we have deduced barely more than half the letters of the alphabet so far (14, to be exact), we have deciphered the vast majority of the letters in the ciphertext, and in fact, the rest is almost trivial. The cipher “l” obviously turns into the plaintext “C,” and with that, things become clearer still, as shown here:

THIS IS AN EpERCISE IN eREAiING ClwHERS USING vREcUENCg DISTRIeUTION. IT IS EASIEST TO USE THIS METHOD lv gOU INTERCEwT A MESSAGE THAT IS MORE THAN ONE HUNDRED CHARACTERS IN LENGTH, eECAUSE IN THAT CASE, THE SAMwLE SluE IS LARGE ENOUGH SO THAT THE CHARACTER DISTRIeUTION IS MORE LliELg TO CLOSELg RESEMeLE THAT Ov A TgwlCAL MESSAGE xRITTEN IN ENGLISH. xE HAoE MADE gOUR TASi EASIER IN THIS EpAMwLE eg SEwARATING THE xORDS IN THE MESSAGE.

Rather than go through the rest of the letters step-by-step, let's look at the message in its entirety:

THIS IS AN EXERCISE IN BREAKING CIPHERS USING FREQUENCY DISTRIBUTION. IT IS EASIEST TO USE THIS METHOD IF YOU INTERCEPT A MESSAGE THAT IS MORE THAN ONE HUNDRED CHARACTERS IN LENGTH, BECAUSE IN THAT CASE, THE SAMPLE SIZE IS LARGE ENOUGH SO THAT THE CHARACTER DISTRIBUTION IS MORE LIKELY TO CLOSELY RESEMBLE THAT OF A TYPICAL MESSAGE WRITTEN IN ENGLISH. WE HAVE MADE YOUR TASK EASIER IN THIS EXAMPLE BY SEPARATING THE WORDS IN THE MESSAGE.