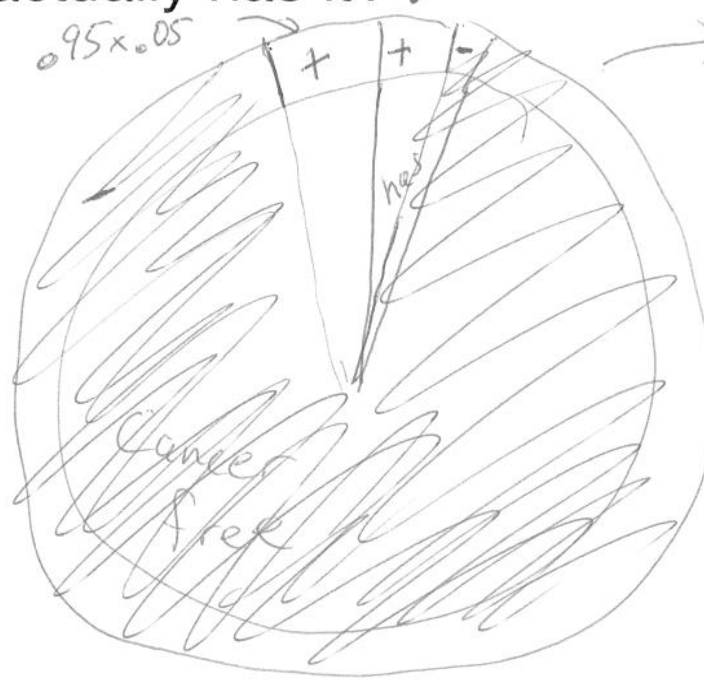
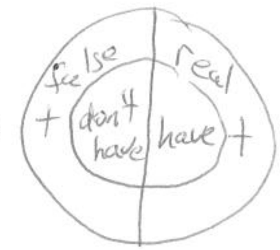
Say that there is a disease which only about 5% of the population has and that there is a test for this disease which is roughly 95% accurate (that is someone who has the disease will test positive 95% of the time and negative 5% of the time, while someone who does not have the disease will test negative 95% of the time and positive 5% of the time).

Given that a patient tests positive for the disease, what is the probability that he or she actually has it? $.05 \times .95$





Is the answer?

6 A) 95%
3 B) 90%
0 C) 75%
7 D) 50%
2 E) Don't know/care

A simple test for monoalphabetic substitution

English:	MISSISSIPPI
Monoalphabetic:	RDFFDFFDOOD
Vigenere :	PQJLLAJBSXZ ← N=11 letters in text

N-1 pairs

In English or monoalphabetic encrypted text we observe:

$$p_{AA} + p_{BB} + p_{CC} + \cdots + p_{ZZ} \approx .027$$

While in polyalphabetic cyphertext we should observe:

$$P(\alpha\alpha \text{ occurrs in random cyphertext }) = \frac{1}{26} \approx .038$$

We should note (of course) that this only works for reasonably large amounts of text.

Look at cyphertext # double letters next to each other = M

if there are N letters in text $\frac{M}{N-1} \simeq .027$ or $.038$

If the cyphertext was obtained from a polyalphabetic cipher then the index of coincidence can also be used to estimate the period of the cipher.

Let p be the period of the cyphertext and place the letters of the cyphertext into groups of p so that the letters in the i^{th} position of the groups are all encrypted with the same key.

- Let $M_\alpha^{(i)}$ equal the number of occurrences of the letter α that appears in the i^{th} positions in the groups.
- If there are M groups of p, then $\sum_{\alpha=A}^{Z} M_\alpha^{(i)} = M$ = # of letters in the i^{th} column
- We also have $N = Mp$
- Also we can estimate that $M_\alpha^{(i)} \approx M p_{\sigma(\alpha)}$ (again for some permutation for the alphabet σ)

$$
M\left\{\begin{array}{l}
i = \overbrace{1\ 2\ 3 \cdots p}^{p} \\
\text{T H I S I} \\
\text{S M Y P L} \\
\text{A I N T E} \\
\text{X T O M O} \\
\text{A R E} \\
\end{array}\right.
$$

Now, we calculate that

pairs of equal letters in same column

pairs of equal letters in different columns

$$
\begin{aligned}
2D_c &= \sum_{i=1}^{p} \sum_{\alpha=A}^{Z} M_\alpha^{(i)}(M_\alpha^{(i)} - 1) + 2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} \sum_{\alpha=A}^{Z} M_\alpha^{(i)} M_\alpha^{(j)} \\
&\approx M^2 p(.065) - pM + M^2(.038)p(p-1) \\
&= \frac{N^2}{p}(.027) - N + N^2(.038)
\end{aligned}
$$

The index of coincidence is defined as

$$I_c = \frac{\text{number of pairs of equal letters in ciphertext}}{\text{the total number of pairs of letters}}$$

not necessarily next to each other

That is if we set

- N_α = the number of occurrences of the letter α in the cyphertext
-
$$D_c = \sum_{\alpha=A}^{Z} \binom{N_\alpha}{2}$$

  D_c represents the number of pairs of equal letters in the cyphertext.
- then $I_c = \frac{D_c}{\binom{N}{2}}$
- where N = the number of letters in the cyphertext

ABCDE

N = # of letters

$\binom{N}{2} = \frac{N(N-1)}{2}$ = # of pairs of letters

The index of coincidence is invariant under monoalphabetic cyphers and we estimate under this condition that $N_\alpha = N * p_{\sigma(\alpha)}$ for some permutation of the alphabet σ and so

$P_{\sigma(\alpha)}$ = prob that α appears in cyphertext

$$\begin{aligned} I_c &= \frac{\sum_{\alpha=A}^{Z}(N_\alpha^2 - N_\alpha)\cancel{/2}}{N(N-1)\cancel{/2}} \\ &\approx \frac{N^2(\sum_{\alpha=A}^{Z} p_\alpha^2) - N}{N(N-1)} \\ &= \frac{N(.065) - 1}{N-1} \\ &\approx .065 \end{aligned}$$

Index of coincidence of English OR Monoalphabetic substitution.

Note that because $I_c = \frac{D_c}{\binom{N}{2}} = D_c/(N(N-1)/2)$, we have that

$$2D_c = N(N-1)I_c.$$

And we just derived that

$$2D_c \approx \frac{N^2}{p}(.027) - N + N^2(.038)$$

Therefore,

$$N(N-1)I_c \approx \frac{N^2}{p}(.027) - N + N^2(.038)$$

$$(N-1)I_c \approx \frac{N}{p}(.027) - 1 + N(.038)$$

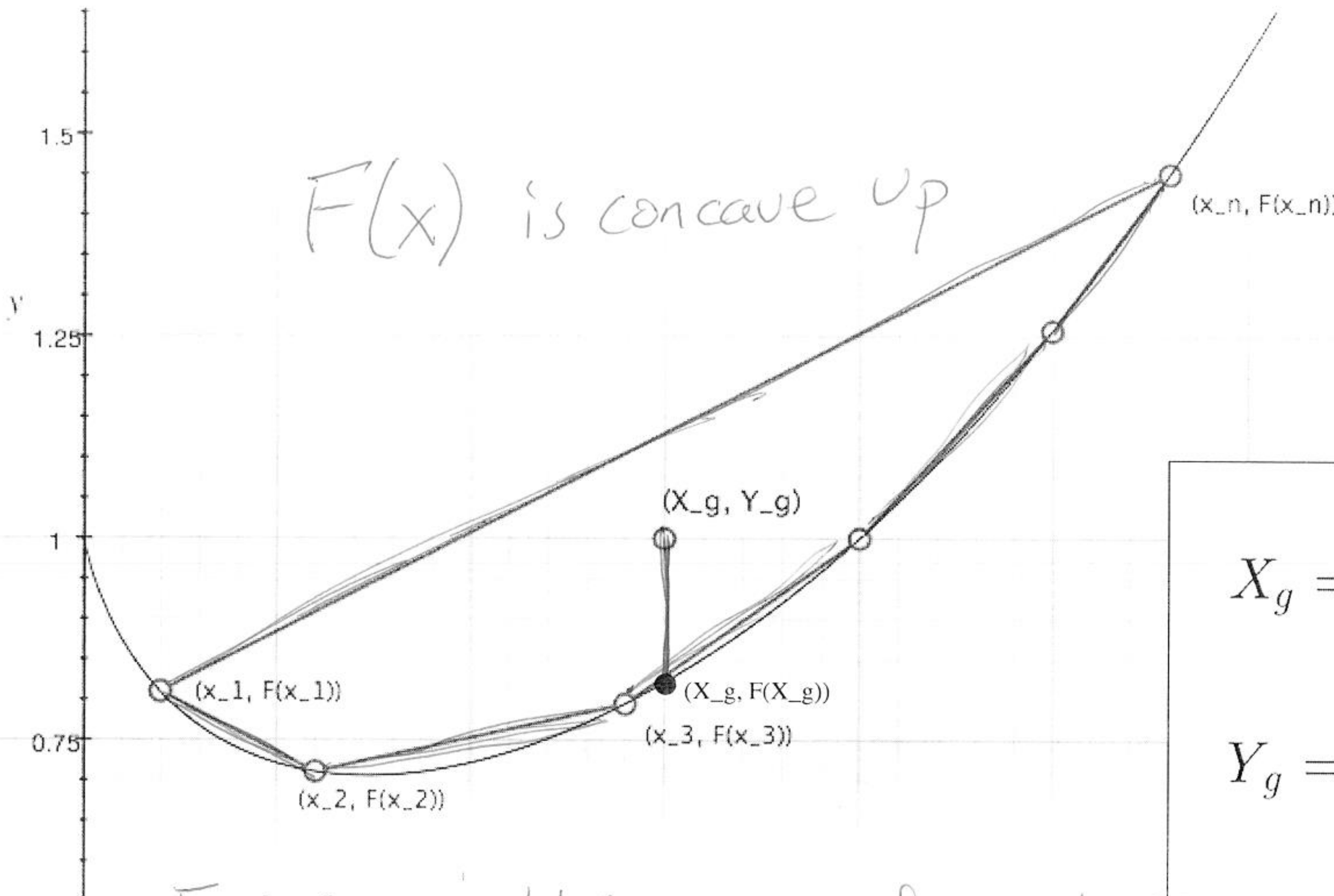
$$(N-1)I_c + 1 \approx \frac{N}{p}(.027) + N(.038)$$

$$(N-1)I_c + 1 - N(.038) \approx \frac{N}{p}(.027)$$

$$p((N-1)I_c + 1 - N(.038)) \approx N(.027)$$

$$p \approx \frac{N(.027)}{(N-1)I_c + 1 - N(.038)}$$

formula for p interms of N and the index of coincidence.



$$m_i \geq 0$$

$$\sum_{i=1}^{n} m_i = 1$$

$$X_g = \sum_{i=1}^{n} m_i x_i$$

$$Y_g = \sum_{i=1}^{n} m_i F(x_i)$$

Fact 1: weighted average of points $(x_i, F(x_i))$ lies inside of convex polygon.

Fact 2: the point which is directly below the weighted average is outside the polygon (or on edge)

$$F\left(\sum_{i=1}^{n} m_i x_i\right) = F(X_g) \leq Y_g = \sum_{i=1}^{n} m_i F(x_i)$$

$$= \sum_{i=1}^{n} m_i y_i$$

$m_i =$ a set of probabilities q_i

where $\sum_{i=1}^{n} q_i = 1$

take another set of probabilities

$p_1, p_2, \ldots, p_n$

Set $\quad x_i = p_i/q_i \quad m_i = q_i$

$$F(x) = x \log_b(x)$$

$$F'(x) = \log_b(x) + x \left(\frac{\ln(x)}{\ln(b)}\right)'$$

$$= \log_b(x) + \cancel{x} \cdot \frac{1}{\cancel{x}\ln(b)}$$

$$F''(x) = \frac{1}{x\ln(b)} \quad \text{concave up!}$$

$$F\left(\sum_{i=1}^{n} \cancel{q_i}\left(\frac{p_i}{\cancel{q_i}}\right)\right) \le \sum_{i=1}^{n} q_i F(p_i/q_i)$$

$$\|$$

$$0$$

$$= \sum_{i=1}^{n} \cancel{q_i}\left(\frac{p_i}{\cancel{q_i}}\right) \log(p_i/q_i)$$

$$= \sum_{i=1}^{n} p_i(\log p_i - \log q_i)$$

Conclusion

$$\sum_{i=1}^{n} p_i \log q_i \leq \sum_{i=1}^{n} p_i \log p_i$$