Data Compression

Past Paper Example

- The frequency of a '0' signal to a binary channel is 0.6. If the '0' is the input, then '0' is the output with probability 0.8.
- If '1' is the input, then output is '1' with probability 0.9.
- Calculate the probability that the input is 1, given that the output is 1.

Past Paper Example

- Let IU be the event that the input is '1' (i.e. I for Input, U rather than O for '1'.)
- Let OU be the event that the output is '1' (i.e. O for output.)
- Let IZ be the event that the input is '0' (i.e. Z for Zero.)
- Let OZ be the event that the output is '0'

- $P(IZ) = 0.6$ Necessarily $P(IU) = 0.4$
- $P(OZ|IZ) = 0.8$ Necessarily $P(OU|IZ) = 0.2$
- $P(OU|IU) = 0.9$ Necessarily $P(OZ|IU) = 0.1$

Example: Part 1

Calculate the probability that the output is '1'.

- $P(OU) = P(OU\ capIU) + P(OU\ capIZ)$
- $P(OU) = [P(OU|IU) \times P(IU)] + [P(OU|IZ) \times P(IZ)]$
- $P(OU) = [0.9 \times 0.4] + [0.2 \times 0.6]$
- $P(OU) = [0.36] + [0.12] = \mathbf{0.48}$

Necessarily $P(OZ) = 0.52$.

Example: Part 2

Calculate the probability that the output is '0' (from first principles).

- $P(OZ) = P(OZ\ capIU) + P(OZ\ capIZ)$
- $P(OZ) = [P(OZ|IU) \times P(IU)] + [P(OZ|IZ) \times P(IZ)]$
- $P(OZ) = [0.1 \times 0.4] + [0.8 \times 0.6]$
- $P(OZ) = [0.04] + [0.48] = \mathbf{0.52}$

Example: Part 3

Compute $P(IU|OU)$

- Use Bayes's Theorem to solve this.

$$P(IU|OU) = \frac{P(OU|IU) \times P(IU)}{P(OU)}$$

- $P(OU|IU) \times P(IU) = 0.9 \times 0.4 = 0.36$
- From before $P(OU) = 0.48$
- Therefore $P(IU|OU) = 0.36/0.48 = \mathbf{0.75}$
- Necessarily $P(IZ|OU) = 0.25$

Example: Part 4

Compute $P(IZ|OZ)$

- Use Bayes's Theorem to solve this.

$$P(IZ|OZ) = \frac{P(OZ|IZ) \times P(IZ)}{P(OZ)}$$

- $P(OZ|IZ) \times P(IZ) = 0.8 \times 0.6 = 0.48$
- From before $P(OZ) = 0.52$
- Therefore $P(IZ|OZ) = 0.48/0.52 = \mathbf{0.923}$
- Necessarily $P(IU|OZ) = 0.077$

- The specifications for the length of a component is a minimum of 9.90 mm and a maximum of 10.44mm.
- A batch of parts is produced that is normally distributed with a a mean of 10.20mm and a standard deviation of 0.20mm.
- Each part costs \$10 to produce. Those that are too short or too long have to be scrapped, or shortened, at a further cost of \$8.
- Compute the percentage of the parts which are (i) Undersize (ii) Oversize.
- Compute the expecte cost of producing 10000 parts.
- Suppose we are able to adjust the processing method such that the mean is halfway between the upper and lower specification.

- The number of undersize parts is therefore 670.
- The number of oversize parts is therefore 1150.
- The number of parts that don't meet specification is therefore 1820.
- The additional cost is $1820 \times 8 = 14,560$

If the mean is changed to 10.17mm, then the symmetry of the normal distribution will mean that the same number of items are too short as too long. To compute this proportion

$$z = \frac{10.44 - 10.17}{0.2} = 1.35$$

- From the tables $P(Z \geq 1.35) = 0.0885$
- 885 items will be too short, 885 items will be too long.

Normal approximation of the binomial distribution

The sample size n should be greater than 30. p must not be too large or too small.

- Normal mean $\mu = np$
- standard deviation $\sigma = \sqrt{np(1-p)}$

Normal approximation of the binomial distribution

Bolts are manufactured by a machine and it is known than approximately 20% are outside certian tolerance limits.
If a random sample of 200 is taken, find the probability that that more than 50 items will be outside the limits.
The normal mean is $200 \times 0.2 = 40$.
The normal standard deviation

$$\sigma = \sqrt{np(1-p)} = \sqrt{200 \times 0.2 \times 0.8} = \sqrt{32}$$

Continuous Uniform Distribution

The probability density function of the continuous uniform distribution, with parameters a and b is given as

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

Continuous Uniform Distribution

The cumulative distribution function of the continuous uniform distribution, with parameters a and b is given as

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{for } x \leq a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b] \\ 1 & \text{for } x \geq b \end{cases}$$

Continuous Uniform Distribution

$$P(L \leq X \leq U) = \frac{U-a}{b-a} - \frac{L-a}{b-a} = \frac{U-L}{b-a}$$

Inference on the Variances of Two Normal Populations

The test statistic used here is based on the distribution.
If s_1^2 and s_2^2 are the sample variances drawn randomly from the two populations and n_1 and n_2 are the two sample sizes, respectively, then the test statistic that can be used to test the equality of the population variances is:

$$F = \frac{s_1^2}{S_2^2}$$

The test statistic follows the F distribution with $(n_1 - 1)$ degrees of freedom in the numerator and $(n_2 - 1)$ degrees of freedom in the denominator.

Quantiles

Recall
The quantile function is the inverse of the cumulative distribution function. The p-quantile is the value with the property that there is probability p of getting a value less than or equal to it. The median is by definition the 50% quantile.
Theoretical quantiles are commonly used for the calculation of confidence intervals and for power calculations in connection with designing and dimensioning experiments.

Example

Ten soldiers visit the rifle range on two different weeks. The first week their scores are:
$\{67, 24, 57, 55, 63, 54, 56, 68, 33, 43\}$
The second week they score $\{70, 38, 58, 58, 56, 67, 68, 77, 42, 38\}$
Give a 95% confidence interval for the improvement in scores from week one to week two.

This is a case of paired samples, for the scores are repeated observations for each soldier, and there is good reason to think that the soldiers will differ from each other in their shooting skill. So we work with the individual differences between the scores. We shall have to assume that the pairwise differences are a random sample from a normal distribution.
The differences are:
$\{3, 14, 1, 3, -7, 13, 12, 9, 9, -5\}$

Effectively we now have a single sample of size 10, and want a 95% confidence interval for the mean of the population from which these differences are drawn. For this we use a Student's t interval. The sample mean of the differences is 5.2, and $s^2 = 54.84$. So $s = 7.41$, and the 95% t interval for the difference in the means is
$5.2 - 2.26(7.41)/\sqrt{10}, 5.2 + 2.26(7.41)/\sqrt{10} = (.0.1, 10.5)$.

Information entropy is often used as a preliminary test for randomness. Generally speaking, random data will have a high level of information entropy, and a low level of information entropy is a good indicator that the data isn't random. (A low level of entropy isn't definitive proof that the data isn't random, but it means you should be suspcious and submit the generator to further tests.)
However, the converse relation doesn't hold, meaning a high degree of information entropy is no guarantee of randomness. For example, a compressed file (e.g., a ZIP file) has a high level of information entropy, but is in fact highly structured, and it will fail many other tests for randomness. Hence, you have to be a little careful using information entropy as a metric for randomness. To get meaningful results, you really need to combine it with other tests.

In information theory, entropy is a measure of the uncertainty associated with a random variable. The term by itself in this context usually refers to the Shannon entropy, which quantifies, in the sense of an expected value, the information contained in a message, usually in units such as bits.
Equivalently, the Shannon entropy is a measure of the average information content one is missing when one does not know the value of the random variable. The concept was introduced by Claude E. Shannon in his 1948 paper ”A Mathematical Theory of Communication”.
A Source X delivers a message among a set of M possible messages x_i , $i = 1, \ldots, M$ with fixed probabilities $p_i = p(x_i)$, $i = 1, \ldots, M$ (pre-defined probabilities). Messages can be seen as events.

Example 1:
Binary source with probabilities p and 1 p
H(X) = p log p (1 p) log(1 p)
Example 2:
M equiprobable messages:
H(x) = logM

Conditional Entropy

Is always lower than a priori uncertainty on X $H(X|Y) < H(X)$

Mutual Information

Mutual information defined as I (X;Y) = H(X) H(X—Y)

- Represents by how much uncertainty decreases on average by knowledge of Y
- Represents information gained on X by observing Y

The t distribution is the appropriate basis for determining the standardized test statistic when the sampling distribution of the mean is normally distributed but s is not known. The sampling distribution can be assumed to be normal either because the population is normal or because the sample is large enough to invoke the central limit theorem.

The t distribution is required when the sample is small ($n < 30$). For larger samples, normal approximation can be used. For the critical value approach, the procedure is identical to that described in Section 10.3 for the normal distribution, except for the use of t instead of z as the test statistic.

The intelligence quotient (IQ) of 36 randomly chosen students was measured. Their average IQ was 109.9 with a variance of 324. The average IQ of the population as a whole is 100.

1. Calculate the p-value for the test of the hypothesis that on average students are as intelligent as the population as a whole against the alternative that on average students are more intelligent.
2. Can we conclude at a significance level of 1% that students are on average more intelligent than the population as a whole?
3. Calculate a 95% confidence interval for the mean IQ of all students.

$$Z_{Test} = \frac{X-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{109.9-100}{\frac{18}{\sqrt{36}}} = \frac{9.9}{3} = 3.3$$

$$p.value = P(Z \geq Z_{Test}) = P(Z \geq 3.3) = 0.00048$$

- $\bar{X} \pm t_{1-\alpha/2,\nu} S.E.(\bar{X})$
- $\nu = 1.96$
- $t_{1-\alpha/2,\nu} = 1.96$
- $109.9 \pm (1.96 \times 3) = [104.02, 115.79]$

For a particular Java assembler interface, the operand stack size has the following probabilities:

Stack Size	0	1	2	3	4
Probability	.15	.05	.10	.20	.50

- Calculate the expected stack size.
- Calculate the variance of the stack size.