Lecture Two

Randomness

4. Statistical Tests

Every binary sequence should have about the same number of ones and zeros, about half the runs (sequences of the same bit) should be of length one, one quarter of length two, one eighth of length three, and so on. They should not be compressible. The distribution of run lengths for zeros and ones should be the same. These properties can be empirically measured and then compared to statistical expectations using a **chi-square test**.

In this section we present some tests designed to measure the quality of a generator purported to be a random bit generator. **While it is impossible to give a mathematical proof that a generator is indeed a random bit generator**, the tests described here help detect certain kinds of weaknesses the generator may have. This is accomplished by taking a sample output sequence of the generator and subjecting it to various statistical tests. Each statistical test determines whether the sequence possesses a certain attribute that a truly random sequence would be likely to exhibit; the conclusion of each test is not definite, but rather probabilistic. An example of such an attribute is that the sequence should have roughly the same number of 0s as 1s. If the sequence is deemed to have failed any one of the statistical tests, the generator may be rejected as being non-random; alternatively, the generator may be subjected to further testing. On the other hand, if the sequence passes all of the statistical tests, the generator is accepted as being random. More precisely, the term "accepted' should be replaced by "not rejected, since passing the tests merely provides probabilistic evidence that the generator

produces sequences which have certain characteristics of random sequences.

4.1 Normal and Chi-square Distributions

The normal and χ^2 distributions are widely used in statistical applications.

Definition(4.1): If the result X of an experiment can be any real number, then X is said to be a **continuous random variable**.

Definition(4.2): A **probability density function** of a continuous random variable X is a function *f*(x) which can be integrated and satisfies:

i. $f(x) \geq 0, \forall x \in R.$

ii. $\int_{-\infty}^{\infty} f(x)dx = 1.$

iii. $\forall a,b \in R, P(a<X<b) = \int_{-\infty}^{\infty} f(x)dx.$

4.1.1 The Normal Distribution

The **normal distribution** arises in practice when a large number of independent random variables having the same mean and variance are summed.

Definition(4.3): A (**continuous**) random variable X has a normal distribution with mean μ and variance σ^2 if its probability density function is defined by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty$$

Remark(4.1): X is said to be $N(\mu,\sigma^2)$. If X is N(0,1), then X is said to have a standard normal distribution.

A graph of the N(0,1) distribution is given in Figure (4.1). The graph is symmetric about the vertical axis, and hence P(X>x)=P(X<-x) for any x.

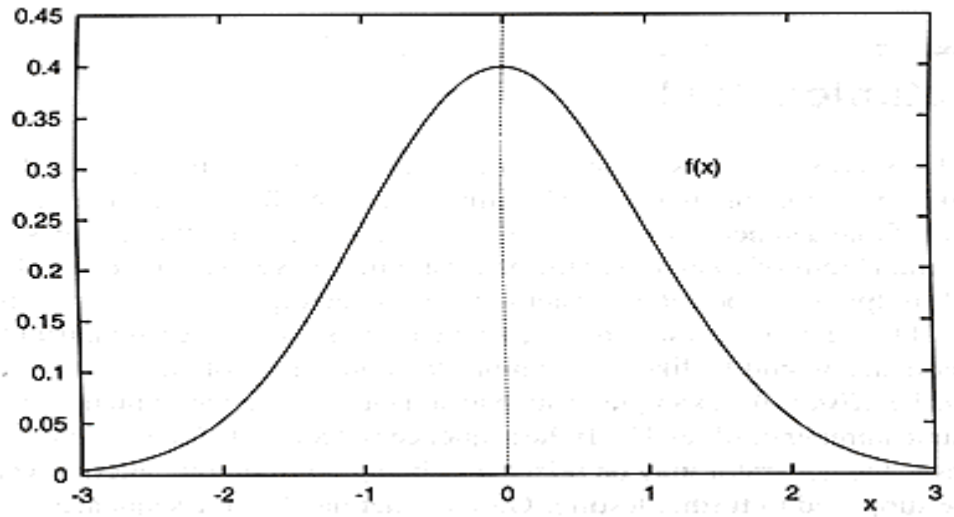

Figure (4.1) The normal distribution N(0,1).

Table (4.1) gives some percentiles for the standard normal distribution. For example, the entry (α=0.05,x=1.6449) means that if X is N(0,1), then X exceeds 1.6449 about 5% of the time.

Table (4.1) selected percentiles of the standard normal distribution. if X is a random variable having a standard normal distribution, then P(X>x)=α.

α	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
x	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902	3.2905

The following remark can be used to reduce questions about a normal distribution to questions about the standard normal distribution.

Remark(4.2): If the random variable X is N(μ,σ^2), then the random variable Z=(X-μ)/σ is N(0,1).

4.1.2 The χ^2 Distribution

The χ^2 (**chi-square**) **distribution** can be used to compare the goodness-of-fit of the observed frequencies of events to their expected

frequencies under a hypothesized distribution. The χ^2 distribution with υ degrees of freedom arises in practice when the squares of υ independent random variables having standard normal distributions are summed.

Definition(4.4): A (**continuous**) random variable X has a χ^2 (chi-square) distribution with υ degrees of freedom if its probability density function is defined by**:**

$$f(x)=\begin{cases}\dfrac{1}{\Gamma(\nu/2)2^{\nu/2}}x^{\nu/2}e^{-x/2}, & 0\le x<\infty\\ 0 & ,x<0\end{cases}$$

where Γ is the gamma function. The mean and variance of this distribution are $\mu = \upsilon$, and $\sigma^2 = 2\upsilon$.

A graph of χ^2 distribution with υ=7 degrees of freedom is given in Figure (4.2).

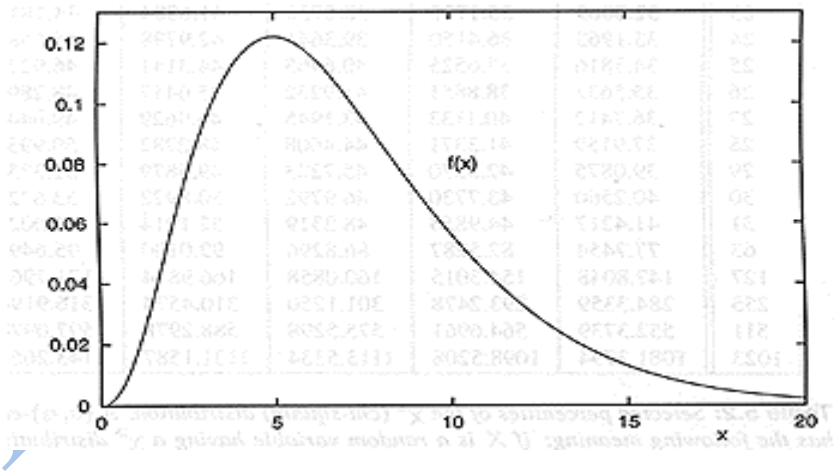

Figure (4.1) The χ^2 distribution (chi-square) with υ=7 degrees of freedom.

Table (4.2) gives some percentiles of the χ^2 distribution for various degrees of freedom degree. For example, the entry υ=5 and column α=0.05 is χ^2=11.0705; this means that if X has χ^2 distribution with 5 degrees of freedom, then X exceeds 11.0705 about 5% of the time.

Table (4.2) selected percentiles of the χ^2 (chi-square) distribution. A(υ,α)-entry of x in the table has the following meaning: If X is a random variable having a χ^2 distribution with υ degrees of freedom, then $P(X>x)=\alpha$.

	α					
υ	0.100	0.050	0.025	0.010	0.005	0.001
1	2.7055	3.8415	5.0239	6.6349	7.8794	10.8276
2	4.6052	5.9915	7.3778	9.2103	10.5966	13.8155
3	6.2514	7.8147	9.3484	11.3449	12.8382	16.2662
4	7.7794	9.4877	11.1433	13.2767	14.8603	18.4668
5	9.2364	11.0705	12.8325	15.0863	16.7496	20.5150
6	10.6446	12.5916	14.4494	16.8119	18.5476	22.4577
7	12.0170	14.0671	16.0128	18.4753	20.2777	24.3219
8	13.3616	15.5073	17.5345	20.0902	21.9550	26.1245
9	14.6837	16.9190	19.0228	21.6660	23.5894	27.8772
10	15.9872	18.3070	20.4832	23.2093	25.1882	29.5883
11	17.2750	19.6751	21.9200	24.7250	26.7568	31.2641
12	18.5493	21.0261	23.3367	26.2170	28.2995	32.9095
13	19.8119	22.3620	24.7356	27.6882	29.8195	34.5282
14	21.0641	23.6848	26.1189	29.1412	31.3193	36.1233
15	22.3071	24.9958	27.4884	30.5779	32.8013	37.6973
16	23.5418	26.2962	28.8454	31.9999	34.2672	39.2524
17	24.7690	27.5871	30.1910	33.4087	35.7185	40.7902
18	25.9894	28.8693	31.5264	34.8053	37.1565	42.3124
19	27.2036	30.1435	32.8523	36.1909	38.5823	43.8202
20	28.4120	31.4104	34.1696	37.5662	39.9968	45.3147
21	29.6151	32.6706	35.4789	38.9322	41.4011	46.7970
22	30.8133	33.9244	36.7807	40.2894	42.7957	48.2679
23	32.0069	35.1725	38.0756	41.6384	44.1813	49.7282
24	33.1962	36.4150	39.3641	42.9798	45.5585	51.1786
25	34.3816	37.6525	40.6465	44.3141	46.9279	52.6197
26	35.5632	38.8851	41.9232	45.6417	48.2899	54.0520
27	36.7412	40.1133	43.1945	46.9629	49.6449	55.4760
28	37.9159	41.3371	44.4608	48.2782	50.9934	56.8923
29	39.0875	42.5570	45.7223	49.5879	52.3356	58.3012
30	40.2560	43.7730	46.9792	50.8922	53.6720	59.7031
31	41.4217	44.9853	48.2319	52.1914	55.0027	61.0983
63	77.7454	82.5287	86.8296	92.0100	95.6493	103.4424
127	147.8048	154.3015	160.0858	166.9874	171.7961	181.9930
255	284.3359	293.2478	301.1250	310.4574	316.9194	330.5197
511	552.3739	564.6961	575.5298	588.2978	597.0978	615.5149
1023	1081.3794	1098.5208	1113.5334	1131.1587	1143.2653	1168.4972

The following remark relates the normal distribution to the χ^2 distribution.

Remark(4.3): If the random variable X is $N(\mu,\sigma^2)$, $\sigma^2>0$, then the random variable $Z=(X-\mu)^2/\sigma^2$ has a χ^2 distribution with 1 degree of freedom. In particular, if X is N(0,1), then $Z=X^2$ has a χ^2 distribution with 1 degree of freedom.

4.2 Hypothesis Testing

Definition (4.5): A **statistical hypothesis**, denoted H_0, is an assertion about a distribution of one or more random variables.

A test of a statistical hypothesis is a procedure, based upon the observed values of the random variables, that leads to the acceptance or rejection of the hypothesis H_0. The test only provides a measure of the strength of the evidence provided by the data against the hypothesis; hence, the conclusion of the test is not definite, but rather probabilistic.

Definition (4.6): The significance level of the test of a statistical hypothesis H_0 is the probability of rejecting H_0 when it is true.

In this section, H_0 will be the hypothesis that a given binary sequence was produced by a random bit generator. If the significance level cat of a test of H_0 is too high, then the test may reject sequences that were, in fact, produced by a random bit generator. On the other hand, if the significance level of a test of H_0 is too low, then there is the danger that the test may accept sequences even though they were not produced by a random bit generator. It is, therefore, important that the test be carefully designed to have a significance level that is appropriate for the purpose at hand; a significance level or between 0.001 and 0.05 might be employed in practice. A statistical test is implemented by specifying a statistic on the random sample. Statistics are generally chosen so that they can be efficiently computed, and so that they (approximately) follow an N(0,1) or a χ^2 distribution. The value of the statistic for the sample output sequence is computed and compared with the value expected for a random sequence.

Suppose that a statistic X for a random sequence follows a χ^2 distribution with υ degrees of freedom, and suppose that the statistic can be expected to take on larger values for nonrandom sequences. To achieve a significance level of α, a threshold value x_α is chosen (using Chi square table) so that $Pr(X>x_\alpha)=\alpha$. If the value X_s of the statistic for the sample output sequence satisfies $X_s>x_\alpha$, then the sequence fails the test; otherwise, it passes the test.