Goodness of fit test (적합도 검정)

1. χ^2 test

1) Definition of χ^2 distribution

When each of random variables $Z_1,\ldots,Z_k$ follows a standard normal distribution N(0,1), and they are independent of one another, the distribution of $Z_1^2 + \ldots + Z_k^2$ is said to be χ^2 distribution with a degree of freedom, k.

$$\sim \qquad (1)$$

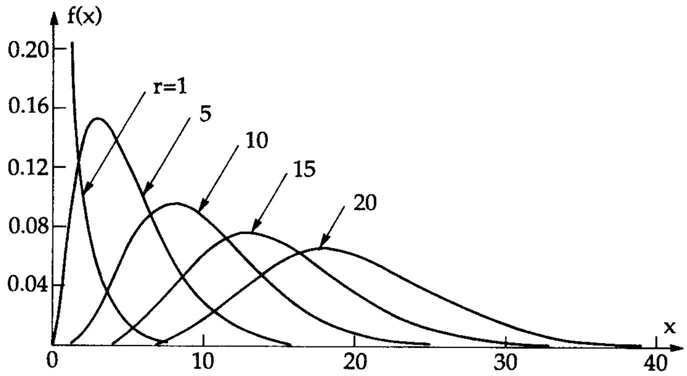


Fig.1 χ^2 distribution with a degree of freedom, r

Probability density function of χ^2 distribution is as follow.

$$, \quad \leq \quad \infty \qquad (2)$$

χ^2 distribution is a kind of gamma distribution with m of r/2 and of 1/2.

$$\overline{\qquad} \qquad (3)$$

where gamma function, ∞

The mean and variance of gamma distribution are , respectively. Therefore, the mean and variance of χ^2 distribution are and , respectively.

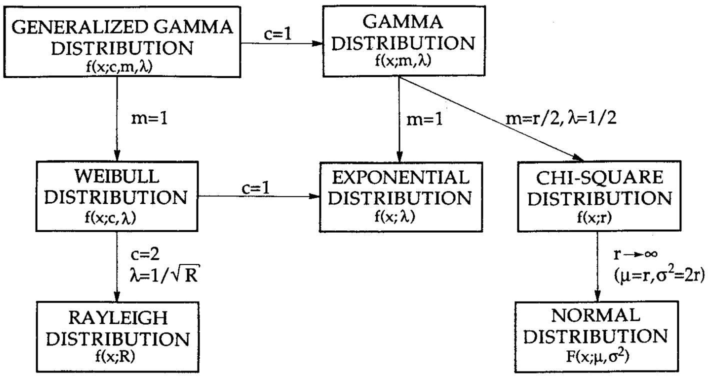
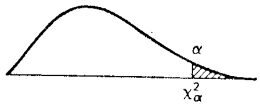


표 6. χ^2 분포표



d.f. \ α	.995	.990	.975	.950	.050	.025	.010	.005
1	392704×10^{-10}	157088×10^{-9}	982069×10^{-9}	393214×10^{-8}	3.84146	5.02389	6.63490	7.87944
2	.0100251	.0201007	.0506356	.102587	5.99147	7.37776	9.21034	10.5966
3	.0717212	.114832	.215795	.351846	7.81473	9.34840	11.3449	12.8381
4	.206990	.297110	.484419	.710721	9.48773	11.1433	13.2767	14.8602
5	.411740	.554300	.831211	1.145476	11.0705	12.8325	15.0863	16.7496
6	.675727	.872085	1.237347	1.63539	12.5916	14.4494	16.8119	18.5476
7	.989265	1.239043	1.68987	2.16735	14.0671	16.0128	18.4753	20.2777
8	1.344419	1.646482	2.17973	2.73264	15.5073	17.5346	20.0902	21.9550
9	1.734926	2.087912	2.70039	3.32511	16.9190	19.0228	21.6660	23.5893
10	2.15585	2.55821	3.24697	3.94030	18.3070	20.4831	23.2093	25.1882
11	2.60321	3.05347	3.81575	4.57481	19.6751	21.9200	24.7250	26.7569
12	3.07382	3.57056	4.40379	5.22603	21.0261	23.3367	26.2170	28.2995
13	3.56503	4.10691	5.00874	5.89186	22.3621	24.7356	27.6883	29.8194
14	4.07468	4.66043	5.62872	6.57063	23.6848	26.1190	29.1413	31.3193
15	4.60094	5.22935	6.26214	7.26094	24.9958	27.4884	30.5779	32.8013
16	5.14224	5.81221	6.90766	7.96164	26.2962	28.8454	31.9999	34.2672
17	5.69724	6.40776	7.56418	8.67176	27.5871	30.1910	33.4087	35.7185
18	6.26481	7.01491	8.23075	9.39046	28.8693	31.5264	34.8053	37.1564
19	6.84398	7.63273	8.90655	10.1170	30.1435	32.8523	36.1908	38.5822
20	7.43386	8.26040	9.59083	10.8508	31.4104	34.1696	37.5662	39.9968
21	8.03366	8.89720	10.28293	11.5913	32.6705	35.4789	38.9321	41.4010
22	8.64272	9.54249	10.9823	12.3380	33.9244	36.7807	40.2894	42.7956
23	9.26042	10.19567	11.6885	13.0905	35.1725	38.0757	41.6384	44.1813
24	9.88623	10.8564	12.4011	13.8484	36.4151	39.3641	42.9798	45.5585
25	10.5197	11.5240	13.1197	14.6114	37.6525	40.6465	44.3141	46.9278
26	11.1603	12.1981	13.8439	15.3791	38.8852	41.9232	45.6417	48.2899
27	11.8076	12.8786	14.5733	16.1513	40.1133	43.1944	46.9630	49.6449
28	12.4613	13.5648	15.3079	16.9279	41.3372	44.4607	48.2782	50.9933
29	13.1211	14.2565	16.0471	17.7083	42.5569	45.7222	49.5879	52.3356
30	13.7867	14.9535	16.7908	18.4926	43.7729	46.9792	50.8922	53.6720
40	20.7065	22.1643	24.4331	26.5093	55.7585	59.3417	63.6907	66.7659
50	27.9907	29.7067	32.3574	34.7642	67.5048	71.4202	76.1539	79.4900
60	35.5346	37.4848	40.4817	43.1879	79.0819	83.2976	88.3794	91.9517
70	43.2752	45.4418	48.7576	51.7393	90.5312	95.0231	100.425	104.215
80	51.1720	53.5400	57.1532	60.3915	101.879	106.629	112.329	116.321
90	59.1963	61.7541	65.6466	69.1260	113.145	118.136	124.116	128.299
100	67.3276	70.0648	74.2219	77.9295	124.342	129.561	135.807	140.169

2) Background theory for goodness of fit test

① Normal approximation to binomial distribution

The mean and variance of a binomial distribution with n times of trials, and p of probability of success, are np and npq (=np(1-p)), respectively. If p is not close to 0 nor to 1 and n is so big that np and n(1-p) are greater than 5, the binomial distribution B(n,p) approaches normal distribution, N(np, np(1-p)).

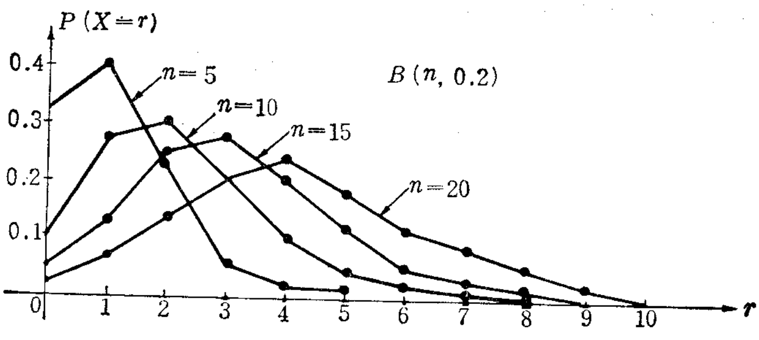


Fig.4 Binomial distribution according to the number of trials

When a random variable X following a binomial distribution satisfies above conditions, we can say that:

(4)

② Multinomial distribution

Let n times of multinomial trials be independently carried out where the probability of the ith event is and the number of the ith events is (i=1,..., k), then

and,

.

This means ~ .

When k=2 so that ~ , below is true.

∵ (5)

Putting equation (1) to equation (5) makes following relation.

$$\frac{\quad}{\quad} \sim \qquad (6)$$

Equation (6) can be generalized as follow.

$$\frac{\quad}{\quad} \sim \qquad (7)$$

Substituting for and for makes

$$= \frac{\quad}{\quad} \qquad (8)$$

Equation (8) enables us to test whether an observed histogram ( ) has a population (모집단) identical to a theoretical distribution ( ). When the level of significance (유의수준: the max. probability that a null hypothesis is rejected (type 1 error)) is the rejection region of the null hypothesis is as follow.

$$\geq \qquad (9)$$

of equation (8) is generally not available in practical problems. Lack of knowledge of the population or its parameters enforces us to estimate the population parameters (모수) from sample statistic (표본통계량). In this case, the degree of freedom of χ^2 distribution reduces as much as the number of population parameters under estimation. If the population is set a normal or lognormal distribution its mean and variance should be estimated, which makes the degree of freedom of χ^2 distribution k-1-2 = k-3.

3) Example of goodness of fit test

The number of rainstorms for past 66 years is as below.

Rainstorms / year	0	1	2	3	4
years	20	23	15	6	2

Test if the number of rainstorms follow Poisson distribution by using a chi-square test with a level of significance of 5%.

Poisson distribution shows the probability that an event with an annual mean of m happens x times in a year as follow.

———

Calculating the mean, m, from the table

— × × × × (times/year)

Calculation procedure is summarized in below table.

No. of storms at station per year	Observed frequency n_i	Theoretical frequency e_i	$(n_i - e_i)^2$	$\frac{(n_i - e_i)^2}{e_i}$
0	20	19.94	0.0036	0.0002
1	23	23.87	0.7569	0.0317
2	15	14.29	0.5041	0.0353
≥ 3	8	7.90	0.0100	0.0013
	66	66.00		0.0685

Because the number of divisions (k) is 4 and parameter under estimation is only m, the degree of chi-square distribution is 2. The chi-square value, 5.99 cannot reject the null hypothesis with the level of significance of 0.05, which means it cannot be said that the annual number of rainstorms does not follow the Poisson distribution.

Ex.1) The number of health insurance claims is known to follow Poisson distribution. An insurance company randomly selected 200 people to investigate the number of claims for past 4 years.

No. of claim	0	1	2	3	4	5	6	7	Total
No. of people	21	54	56	38	23	4	3	1	200

Test by using the data and the level of significance of 2.5% whether there is any reason to say that the number of claims does not follow Poisson distribution.

2. Kolmogorv-Smirnof (K-S) test

1) Test procedure

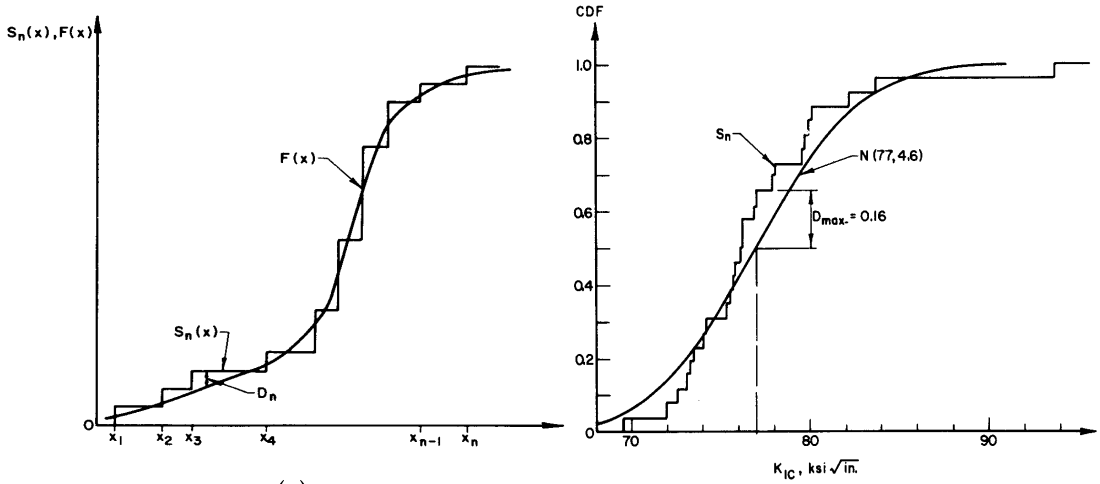
- Arrange the data in order of magnitude so that it seems to be a cumulative distribution.

— ≤

- Compare the theoretical cumulative distribution (F(x)) with the sampling data and get the maximum difference between the two in absolute value, that is = Max

.

- If is less than the null hypothesis cannot be rejected, which means both distributions are not different each other for the level of significance, .



(a) (b)

Fig.5. Comparison of theoretical and observed distributions for K-S test

2) Difference from χ^2 test

For χ^2 test, each division of a sample histogram should have a frequency greater than 5, while K-S test does not have this kind of restriction. Since the sample distributions usually have small frequencies in tail, data in the tail should be bound together to make its frequency higher than 5. This causes the K-S test to more sensitively respond to the difference around a tail than χ^2 test.

Ex.2) Errors in 8 measurements are given below. Apply K-S test to testing whether the measurement error follow a normal distribution of which mean and standard deviation are 0 and 0.1, respectively.

0.07 0.12 -0.06 -0.04 -0.05 0.08 0.04 0.00

Table A.4. Critical Values of D_n^α in the Kolmogorov-Smirnov Test (*After Hoel, 1962*)

n \ α	0.20	0.10	0.05	0.01
5	0.45	0.51	0.56	0.67
10	0.32	0.37	0.41	0.49
15	0.27	0.30	0.34	0.40
20	0.23	0.26	0.29	0.36
25	0.21	0.24	0.27	0.32
30	0.19	0.22	0.24	0.29
35	0.18	0.20	0.23	0.27
40	0.17	0.19	0.21	0.25
45	0.16	0.18	0.20	0.24
50	0.15	0.17	0.19	0.23
>50	$1.07/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

Box 2.26 Worked examples: Kolmogorov–Smirnov test for normal distribution

1 The data are errors in eight observations and it is suspected that they come from a normal distribution with mean equal to zero and standard deviation equal to 0.1.

0.07 0.12 −0.06 −0.04 −0.05 0.08 0.04 0.00

The hypotheses are:

H_0: the distribution is normal with $\mu = 0$, $\sigma = 0.1$
H_1: the data come from some other distribution

The data are ordered and values of the sample cumulative relative frequencies and theoretical values of $F(x_1), \ldots, F(x_n)$ are calculated, using the table of the normal distribution. These values have been tabulated and plotted (Fig. B2.26.1). Note that the cumulative relative frequency of the data is plotted as a step function.

Table B2.26

Error	Cumulative frequency	C.f./$(n + 1)$	z	$\Pr(Z \leqslant z)$
−0.06	1	0.11	−0.6	0.2743
−0.05	2	0.22	−0.5	0.3085
−0.04	3	0.33	−0.4	0.3446
0.00	4	0.44	0.0	0.5000
0.04	5	0.56	0.4	0.6554
0.07	6	0.67	0.7	0.7580
0.08	7	0.78	0.8	0.7881
0.12	8	0.89	1.2	0.8449

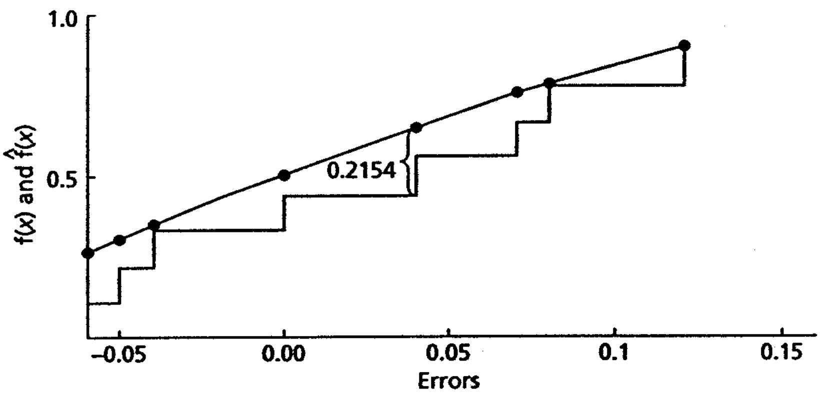


Fig. B2.26.1

continued on p. 124