Lec 14

A population contains a group of elements (e.g. seeds or data) taken into account in a study.

A sample is a subgroup of elements taken randomly from a population in order to represent it. A good sample for seed testing must be representative of the lot.

A variable is a numerical measurement made on a population member, or a sample member. Variables are of two types: discrete and continuous.

A discrete variable is one which is restricted to a number of admitted values only and involves counting (e.g. the number of germinated seeds in the sample).

A continuous variable is one which can represent any value in a given range and involves measurement (e.g., seed moisture).

The expected value is the mean of a population E(x).

The expected value can be estimated by the mean value of sample data:

$$\bar{x} = \frac{\sum x}{N},$$

where N= the number of sample elements and Σ means sum of the values.

A homogeneous population is one in which seed lot values (quality values) are dispersed around expected value of the population within acceptable limits.

The variance is the most important measure of dispersion around the expected value of the population V(x).

The variance of the lot can be estimated by

$$V = \frac{N\sum x^2 - (\sum x)^2}{N(N-1)}.$$

The standard deviation is also a measure of dispersion around the expected value of population; it is the square root of the variance, SD(x)= $\sqrt{V(x)}$.

The range is another measure of dispersion, indicating the maximum difference between the observed values within the lot.

$$R = x_{max} - x_{min}$$

The above parameters of a lot can be estimated by the measured data from a sample. It should be noted that the estimation of lot parameters on the basis of sample data is not reliable unless the lot is genuinely homogeneous (within acceptable limits). A sample taken from a heterogeneous lot is unlikely to be representative. The probability is low that an estimate based on such a sample will be satisfactory .

PROBABILITY DISTRIBUTIONS

A discrete probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable, x.

The probability function f(x) gives the probability for each value of the random variable. It is to be noted that

$$f(x) \geq 0,$$

$$\sum f(x) = 1.$$

The expected value is

$$E(x) = \sum x f(x).$$

The variance is

$$V(x) = \sum \left[x - E(x) \right]^2 f(x).$$

THE MOST IMPORTANT DISCRETE PROBABILITY DISTRIBUTIONS

4.1. The binomial probability distribution

The binomial distribution can be applied when a population consists of two different types of elements (e.g., healthy and not healthy, germinated and not germinated).

Suppose that the probability of showing a special characteristic of any single member of the population is p and therefore q=(1-p) is the probability of not showing it. It is important that p and q are proportions and not percentages. E.g., for germination, this definition implies that p=0 only when no seeds have germinated, and p=1 only when all seedlings have germinated. The p is constant during an experiment.

Assume a random sample of size n from the whole population and denote with x the number of observations having the special characteristic out of n. In this case x is a discrete random variable with possible values 0, 1, 2, ...n which are not equally likely. The binomial probability function f(x) shows the probability of different x values. The n and p values are characterising parameters of the binomial distribution.

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} .$$

Expected value and variance of the binomial distribution are:

$E(x)=np$, $V(x)=npq$, where q=1-p.

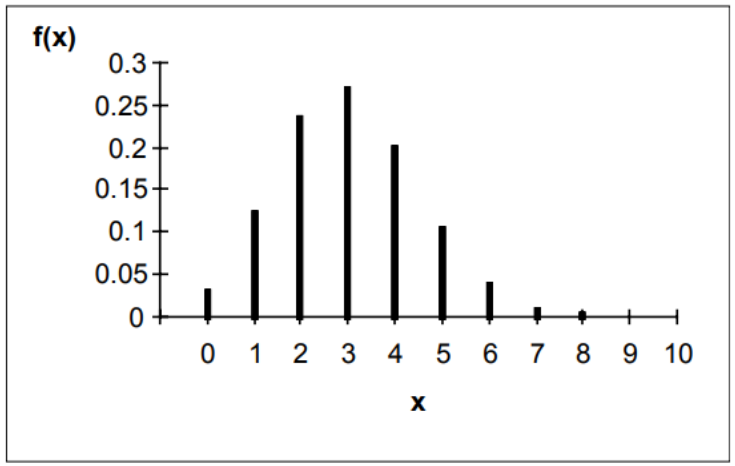

Figure 1

Binomial probability function, n=10 p=0.3.

4.2. The Poisson distribution

The Poisson distribution gets its name after the French mathematician who first studied and applied it. He showed that the Poisson distribution is the marginal case of the binomial distribution when the parameter p tends to zero and simultaneously n tends to infinite. So this distribution can be applied when a population contains only a very small number of members of a special characteristic but a large sample has to be examined, e.g. other seeds by number in a seed lot.

$p \to 0$, $n \to \infty$ and np= constant (denoted with λ).

This distribution is also a discrete one, and gives the probability of x which shows the occurrence of a rare event in a large sample. The possible values of x are 0,1,2,3,4,... (no upper limit).

The probability function f(x) of the parameter λ is given:

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{where } x = 0,1,2,3...$$

x!=x(x-1)...1 product and e≈ 2.7183.. , is an irrational number which is the base of the natural logarithm. It can be proved that the expected value and the variance both equal λ. This value is the single characterising parameter of the Poisson distribution.

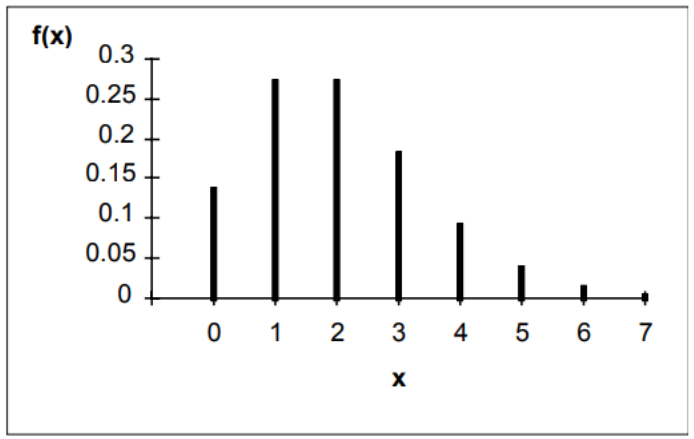

Figure 2

Poisson probability function if λ=2.

The chi-square test is a statistical method used to determine if there is a significant association between categorical variables. It's particularly useful for analyzing categorical data and assessing whether observed frequencies in a contingency table (a type of table used to display the frequency distribution of variables) differ significantly from expected frequencies under the null hypothesis.

There are two main types of chi-square tests:
1.Chi-Square Test of Independence: This test evaluates whether two categorical variables are independent of each other. For example, you might use it to determine if there's an association between gender and voting preference. The test compares the observed frequency of cases in each category of a contingency table with the frequencies we would expect if the variables were independent.

2.Chi-Square Test of Goodness of Fit: This test assesses whether the distribution of a single categorical variable conforms to an expected distribution. For instance, it could be used to check if the number of people falling into different age groups is consistent with what we would expect based on a theoretical distribution.

Methodology:

1. Formulate Hypotheses:

1. Null Hypothesis (H0): Assumes no effect or no association (e.g., the variables are independent, or the observed distribution fits the expected distribution).

2. Alternative Hypothesis (H1): Assumes there is an effect or an association.

2. Calculate the Chi-Square Statistic: The formula is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency, and E_i is the expected frequency for each category.

3. Determine the Degrees of Freedom:

For a test of independence, it is **(number of rows−1)×(number of columns−1)**

For a goodness-of-fit test, it is **number of categories−1**

4. Compare to the Critical Value: Use a chi-square distribution table to find the critical value for the calculated degrees of freedom and significance level. If the calculated chi-square statistic exceeds the critical value, reject the null hypothesis.

Chi-Square Distribution Table is a probability table of selected values of X^2

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

1. Classroom Tasks
2. Assignment Problems