Second course in Statistics. Lecture 19.

Inference of two way table:

- Chi-square test of independence

Inference of one way table:

- Chi-square goodness of fit test

One-way analysis of variance (ANOVA)

Chi-square test of independence

Setting 1:

In example of note 18, a single SRS is drawn from a population, and observations are classified according to two categorical variables having r and k possible values.

H_0 : Column and row variables are independent.

Setting 2:

Independent SRSs are drawn from each of k populations and each observation is classified according to a categorical variable with r possible values.

H_0 : The distributions of row variable are the same for all k populations.

Chi-square test of independence

Example 1: Market researchers know that background music can influence the mood and purchasing behaviour of customers. One study in a supermarket in Northern Ireland compared three treatments: no music, French accordion music, and Italian string music. Under each condition, the researchers recorded the numbers of bottles of French, Italian, and other wine purchased. Here is a two-way table that summarizes the data.

music / wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

Chi-square test of independence

Column percentage

music / wine	None	French	Italian	Total
French	35.71%	52.00%	35.71%	40.74%
Italian	13.10%	1.33%	22.62%	12.76%
Other	51.19%	46.67%	41.67%	46.50%
Total	100%	100%	100%	100%

Chi-square test of independence

Expected cell frequencies under the assumption that wine type sold is independent of music playing.

music / wine	None	French	Italian	Total
French	30 (34.2)	39 (30.6)	30 (34.2)	99
Italian	11 (10.7)	1 (9.6)	19 (10.7)	31
Other	43 (39.1)	35 (34.9)	35 (39.1)	113
Total	84	75	84	243

Chi-square goodness of fit test

Data for n observations on a categorical variable with k possible outcomes are summarized as observed counts, $n_1, n_2, \mathrm{L}\, n_k$ in k cells. We want to compare it with a hypothesized distribution. A null hypothesis specifies probabilities $p_1, p_2, \mathrm{L}\ p_k$ for the possible outcomes.

Example 2: A die is tossed 360 times and the results are shown as follows:

	1	2	3	4	5	6	sum
Observed frequency	49	52	47	71	73	68	360

Test at 5% level of significance whether the die is fair.

Chi-square goodness of fit test

Assumptions

X_i's are iid (from a random sample).

E_i for each class must be larger than 1

No more than 20% E_i's are less than 5.

Hypothesis:

$H_0 : p_i = p_{i0}$ for all values of i

$H_a : p_i \neq p_{i0}$ for some values of i

Examples: $H_0 : X \sim n(\mu, \sigma^2)$ $H_0 : X \sim Bin(n, p)$ H_0: A die is fair

$H_a : X \neq n(\mu, \sigma^2)$ $H_a : X \neq Bin(n, p)$ H_a: A die is not fair

Chi-square goodness of fit test

Test statistic: (commonly used)

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad \text{or} \quad \chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n$$

where O_i is the observed frequency in class i

$E_i = np_{i0}$ is the expected frequency in class i under H_0

Degree of freedom $\upsilon = k - 1 - r$

k is the number of classes

r is the number of estimated population parameter.

Rejection region:

Reject H_0 if $\chi^2 > \chi_\alpha^2(k-1-r)$ at α level of significance.

One-way ANOVA

The statistical methodology for comparing several means is called analysis of variance, or simply ANOVA.

When there is only one way to classify the population of interest, we use one-way ANOVA, to analyze the data.

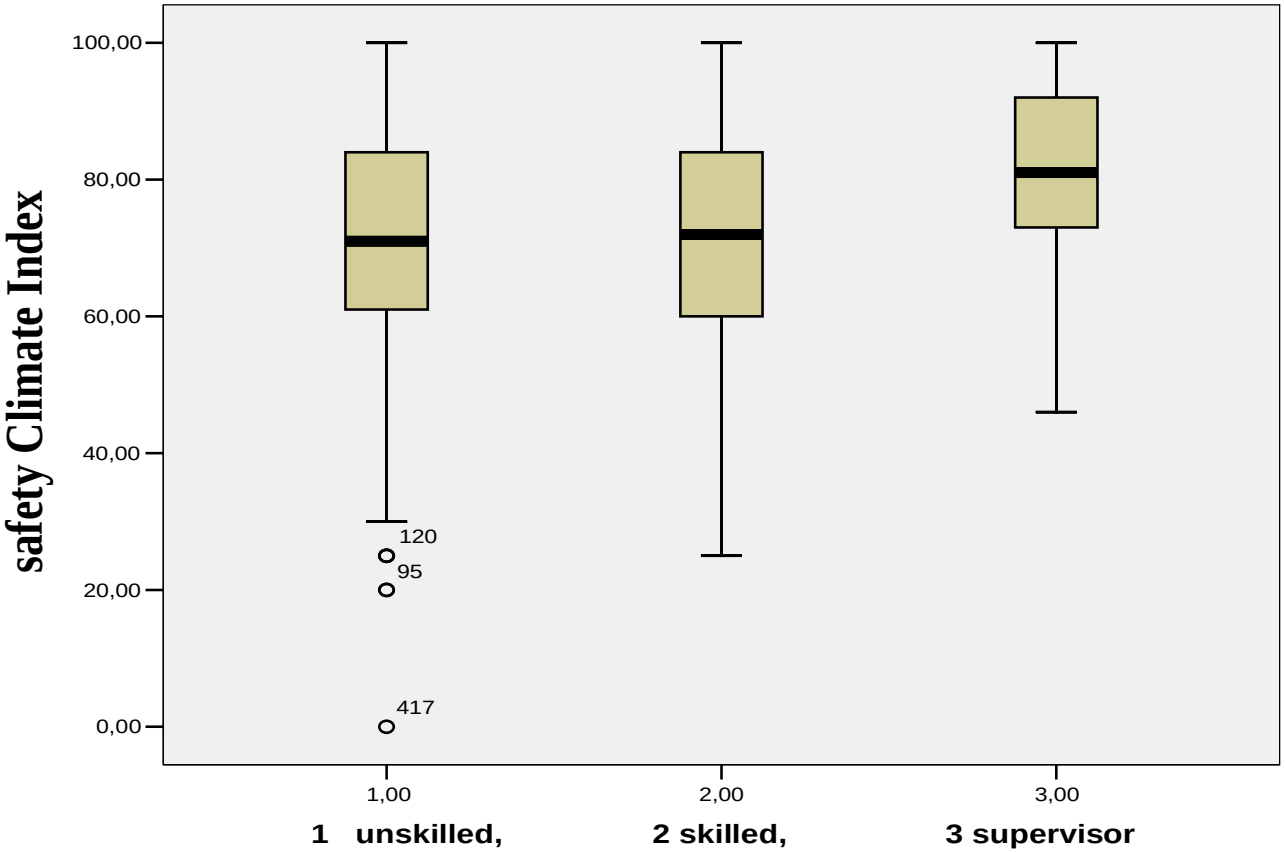
If there is more than one way to classify the populations, we use two-way ANOVA. For example, a mail-order firm might want to compare mailings that offer different discounts and also have different layouts. Will a lower price offered in plain format draw more sales on the average than a higher price offered in a fancy brochure? Analyzing the effect of price and layout together requires two-way ANOVA.

Example 3

In a study of workplace safety, workers were asked to rate the safety of their work environment, and a composite score called the Safety Climate Index (SCI) was calculated. The index is the sum of the responses to 10 different questions about safety. The response for each of these questions is an integer ranging from 0 to 10, so the SCI has values from 0 to 100. The Workers were classified according to job category as unskilled, skilled, and supervisor. Data in an extra excel file. Here is a summary of the data:

Job category	n	$\bar{x}$	s
Unskilled workers	448	70.42	18.27
Skilled workers	91	71.21	18.83
Supervisors	51	80.51	14.58

One-way ANOVA



One-way ANOVA

Assumptions:

- Populations are normal with the same standard deviation.
- If we have unequal standard deviations, we generally try to transform the data so that they are approximately equal.
- Formal tests for equality of standard deviations in several groups lack the robustness against nonnormality (we know the case of two groups).
- ANOVA is not extremely sensitive to unequal standard deviations. If the largest standard deviation is less than twice the smallest standard deviation, we can use ANOVA based on the assumptions of equal variances, and our result will still be approximately correct.

One-way ANOVA

Hypothesis:

H_0 : The population means are all equal. $\mu_1 = \mu_2 = \mathrm{L} = \mu_k$

H_a : They are not equal. $\mu_i \neq \mu_j$ for some $i \neq j$

Remarks:

This alternative could be true because all of the means are different or simply because one of them differs from the rest.

If we reject null hypothesis, we need to perform some further analysis to draw conclusions about which population means differ from which others.

One-way ANOVA

The one-way ANOVA model is

$$x_{ij} = \mu_i + \varepsilon_{ij}$$

for $i = 1, \text{L}\, k$ and $j = 1, \text{L}\, , n_i$. The ε_{ij}'s are random error terms and assumed to be from an $N(0, \sigma^2)$ distribution. The parameters of the model are the population means $\mu_1, \mu_2, \text{L}\, \mu_k$ and standard deviation σ.

Sample size n_i may differ

Standard deviation σ is assumed to be the same in all populations.

One-way ANOVA

Sample statistics:

Groups

1	2	...	j	...	k
Y_{11}	Y_{12}	L	Y_{1j}	L	Y_{1k}
Y_{21}	Y_{22}	L	Y_{2j}	L	Y_{2k}
M	M	M	M	M	M
Y_{i1}	Y_{i2}	L	Y_{ij}	L	Y_{ik}
M	M	M	M	M	M
$Y_{n_1 1}$	$Y_{n_2 2}$	L	$Y_{n_j j}$	L	$Y_{n_k k}$

One-way ANOVA

Sample statistics:,

Sum for each groups: $T_{\cdot j} = \sum_{i=1}^{n_j} Y_{ij}$, $j = 1,2, \mathsf{L}, k$

Sample mean for each group: $\overline{Y}_{\cdot j} = T_{\cdot j} / n_j$ $j = 1,2, \mathsf{L}, k$

Sum of total: $T_{\cdot\cdot} = \sum_{j=1}^{k} T_{\cdot j}$,

Total number of observations: $N = \sum_{j=1}^{k} n_j$,

Overall sample mean: $\overline{Y}_{\cdot\cdot} = T_{\cdot\cdot} / N$.

$$Y_{ij} = \mu + (\mu_j - \mu) + (Y_{ij} - \mu_j) \Rightarrow Y_{ij} - \overline{Y}_{\cdot\cdot} = (\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}) + (Y_{ij} - \overline{Y}_{\cdot j})$$

One-way ANOVA

Sample statistics:

$Y_{ij} - \overline{Y}_{\cdot\cdot}$ deviations of each observation from overall mean

$\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}$ deviations of group mean from overall mean

$Y_{ij} - \overline{Y}_{\cdot j}$ deviations of each observation from group mean.

Square both sides of $Y_{ij} - \overline{Y}_{\cdot\cdot} = \left(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}\right) + \left(Y_{ij} - \overline{Y}_{\cdot j}\right)$ and sum over all i and j, thus

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}\right)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_{\cdot j}\right)^2$$

This equation is known as analysis of variance.

One-way ANOVA

Sample statistics:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij}-\overline{Y}_{\cdot\cdot}\right)^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot}\right)^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij}-\overline{Y}_{\cdot j}\right)^2$$

Sum of squares:

Total sum of squares: $SST=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij}-\overline{Y}_{\cdot\cdot}\right)^2$

Group sum of squares: $SSG=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot}\right)^2$

Sum of squares error: $SSE=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij}-\overline{Y}_{\cdot j}\right)^2$

One-way ANOVA

Degrees of freedom are related to the deviations that are used in the sums of squares.

Total degree of freedom: $DFT = N - 1$

Group degree of freedom: $DFG = k - 1$

Degree of freedom error $DFE = N - k$

$$DFT = DFG + DFE$$

Mean squares:

Group mean square: $MSG = SSG/(k-1)$

The error mean square: $MSE = SSE/(N-k)$

One-way ANOVA

F test statistics

$$F = \frac{MSG}{MSE} \text{ with } \nu_1 = k - 1 \text{ and } \nu_2 = N - k$$

Rejection region: Reject H_0 if F is sufficiently large.

Example 4

We are interested in whether slightly different amounts of residential ceiling insulation have an effect on energy consumption. Data is recorded as the kilowatt hours used by the heating systems of very similar houses in a given month as a function of five levels of ceiling insulation (in inches). Is there reason to believe that at least some of the average energy consumptions for the five levels of ceiling insulation are different? $\alpha = 0.01$

Thickness of ceiling insulation (inches)

4	6	8	10	12
14.4	14.5	13.8	13.0	13.1
14.8	14.1	14.1	13.4	12.8
15.2	14.6	13.7	13.2	12.9
14.3	14.2	13.6		13.2
14.6		14.0		13.3
				12.7