Worksheet – Sections 10.3

Hypothesis test of association asks one of two following questions:

Are two different variables of a population associated with each other? (**Test of independence**)

- H_0: they are independent
- H_a: they are associated

Does a certain variable have the same distribution between two populations? (**Test of homogeneity**)

- H_0: the distributions are the same
- H_a: the distributions are different

A **contingency table** (or **two-way table**) of two variables displays the frequencies of observations for every possible pairing of their values.

Expected frequency of each pairing:

$$E = \frac{(row\ total) \times (column\ total)}{grand\ total}$$

Chi-square statistic measures the discrepancy between the observed frequencies and the expected frequencies:

$$X^2 = sum\ of\ \frac{(O-E)^2}{E}\ over\ all\ possible\ pairings$$

where O and E are the observed frequency and expected frequency of each pairing.

X^2 is approximately *chi-square* distributed with $(k_1 - 1)(k_2 - 1)$ degrees of freedom, where k_1 and k_2 are respectively the numbers of possible values of the first and second variables, if the following conditions are satisfied:

- Simple random sampling.
- Large sample: every expected frequency is at least 5.

p-value = P($X^2 \geq$ observed value of X^2). Reject H_0 if the p-value is lower than the significance level.

StatCrunch to compute p-value: Stat → Tables → Contingency → With Summary

1) A health researcher wants to test if smoking is associated with lung disease. He asked randomly 300 people and recorded each person's smoking status and whether he/she had lung disease. The outcome is summarized as follows:

Smoking status	Lung disease	No lung disease
Smoker	45	105
Non-smoker	15	135

a) Is this a test of homogeneity or a test of independence?

b) State the null hypothesis and alternative hypothesis.

c) Find the expected frequency for each paring of smoking status and disease status.

d) Compute the chi-square statistic X^2.

e) Check if the conditions guaranteeing that X^2 is chi-square distributed are satisfied.

f) Compute the chi-square statistic X^2 and p-value using Statcrunch.

g) At the 5% significance level, would you conclude that smoking is associated with lung disease?

2) A researcher is interested in whether the blood-type distribution is the same on two neighboring islands. A random sample of 300 residents is selected from Island 1, and an independent random sample of 300 residents is selected from Island 2. The outcome is summarized as follows:

Blood type	Island 1	Island 2
A	120	85
B	40	55
AB	15	20
O	125	140

a) Is this a test of homogeneity or a test of independence?

b) State the null hypothesis and alternative hypothesis.

c) Compute the chi-square statistic X^2 and p-value using Statcrunch.

d) At the 5% significance level, would you conclude that the blood-type distributions are the same on the two islands?