Lecture 15 Categorical data and chi-square tests

- Continuous variable: height, weight, gene expression level, lethal dosage of anticancer compound, etc --- ordinal
- Categorical variable: sex, profession, political party, blood type, eye color, phenotype, genotype
- Questions: do smoke cause lung cancer? Do smokers have a high lung cancer rate?
- Do the 4 nucleotides, A, T, G, C, occur equally likely?
Sample space: the set of possible basic outcomes

- To study categorical variables, the first thing is to know what the categories are.
- face of coin: head, tail
- face of a die: 1, 2, 3, 4, 5, 6
- Nucleotide: A, T, G, C
- Sex: male, female
- Blood type: A, B, O, AB

The set of possible outcome of a categorical variable forms a sample space.

When two categorical variables are involved, then the sample space is the set of all possible combinations.
Subjective probability and assumption of independence

- Symmetry: if two outcomes are deemed symmetrical, then they should be assigned with an equal probability.
- Sum of probability is equal to 1.
- If two variables are independent, then you can multiply the probability.
- Statistical questions: can symmetry be assumed? Can independence be assumed?
- Solution: Collect data and conduct a chi-square test.
Examples

• A random sample of 100 nucleotides is obtained. There are 24 A, 21 T, 30 G, 25 C.

• Are the data compatible with the assumption of equal occurrence?

• Suppose G and C are mixed up by error. So we have 24 A, 21 T, 55 G/C. What is the answer?