# MAT 2379, Introduction to Biostatistics

## Chapter 4. Conditional Probability
## 4.1 Definition

Consider again the example with a group of 100 German men of which 70 have blond hair, 77 have blue eyes and 55 have blond hair and blue eyes. Suppose that we select randomly a man in the group of people which have blond hair (i.e. suppose that we *already know* that the selected person has blond hair). What is the probability that this person has also blue eyes?

Because we make our selection in the restricted group of 70 blond hair people and in this group only 55 satisfy the requirement of having blue eyes, the desired probability is

$$P(\text{blue eyes}|\text{blond hair}) = \frac{55}{70} = 0.79$$

This is called **the conditional probability** of the event $B$: "the selected person has blue eyes" given the event $A$: "the selected person has blond hair". This probability shows us that there is a great degree of dependence between the two events.

In general, the formula for calculating the conditional probability is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

The key words that indicate that we are interested in calculated a conditional probability are: **"if"** and **"given that"**.

*Example 1.* In a study of the relationship between health risk and income which was performed on a large group of people living in the same region, it was found that 43% smoke, 70% are stressed and 85% smoke or are stressed. Let $A$ be the event "the person smokes" and $B$ be the event "the person is stressed". Using the addition rule,

$$P(A \cap B) = 0.43 + 0.70 - 0.85 = 0.28$$

The probability that the selected person is stressed, given that this person is smoking is

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.28}{0.43} = 0.651$$

The probability that the selected person is smoking, given that this person is not stressed is

$$P(A|B') = \frac{P(A \cap B')}{P(B')} = \frac{0.15}{0.30} = 0.5$$

since $P(A \cap B') = P(A) - P(A \cap B) = 0.43 - 0.28 = 0.15$ and $P(B') = 1 - P(B) = 0.30$.

**Diagnostic Tests**

A diagnostic test is a test given to detect the presence of some specific medical condition: pregnancy test, HIV test, DOWN syndrome test (on pregnant women), etc.

In a diagnostic test, only one of following 4 things can happen:

1. The subject is a true positive and the test result is positive. No error has been made.

2. The subject is a true negative but the test result is positive. An error has been made, which can result in inconvenience and expense for the subject. This error is called the **false-positive rate** of the test and is denoted by $\alpha$:

$$\alpha = P[\text{test result is positive}|\text{subject is a true negative}]$$

3. The subject is a true positive and the test result is negative. No error has been made.

4. The subject is a true positive but the test result is negative. An error has been made, which is potentially dangerous. This error is called the **false-negative rate** of the test and is denoted by $\beta$:

$$\beta = P[\text{test result is negative}|\text{subject is a true positive}]$$

Some other rates can be calculated:

$$\text{specificity} = P[\text{test result is negative}|\text{subject is a true negative}]$$

$$\text{sensitivity} = P[\text{test result is positive}|\text{subject is a true positive}]$$

$$\text{positive predictive value } (PPV) = P[\text{subject is a true positive}|\text{test result is positive}]$$

$$\text{negative predictive value } (NPV) = P[\text{subject is a true negative}|\text{test result is negative}]$$

*Example 2.* Currently in Ontario, pregnant women (under the age of 35) are asked if they want to take a blood test which may reveal whether their child has the DOWN syndrome, a very serious genetic disease. (The test is strongly recommended for women over 35). This test is not very accurate and therefore women should be informed of the risks that are involved.

Suppose that in a sample of 300 pregnant women who have decided to take the test, 5 obtained a positive test result and 295 obtained a negative test result. Suppose that from the 5 women who obtained a positive test result, 4 were truly carrying children with the DOWN syndrome but 1 woman had a perfectly normal baby. Furthermore, suppose that from the 295 women who obtained a negative test result, 3 were in fact carrying children with the DOWN syndrome. One can draw a table which illustrates the 4 possible situations (in class).
From the table we can estimate the following rates

$$\alpha = P[\text{test } +|\text{true } -] = \frac{1/300}{293/300} = \frac{1}{293} \approx 0.003$$

$$\beta = P[\text{test } -|\text{true } +] = \frac{3/300}{7/300} = \frac{3}{7} \approx 0.429$$

$$\text{specificity} = P[\text{test } -|\text{true } -] = \frac{292/300}{293/300} = \frac{292}{293} \approx 0.997$$

$$\text{sensitivity} = P[\text{test } +|\text{true } +] = \frac{4/300}{7/300} = \frac{4}{7} \approx 0.571$$

$$\text{positive predictive value} = P[\text{true} + | \text{test} +] = \frac{4/300}{5/300} = \frac{4}{5} = 0.8$$

$$\text{negative predictive value} = P[\text{true} - | \text{test} -] = \frac{292/300}{295/300} = \frac{292}{295} = 0.989$$

## 4.2 Multiplication Rule

The multiplication rule helps us to calculate $P(A \cap B)$ using one of the conditional probabilities $P(A|B)$ or $P(B|A)$. The rule says:

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

*Example 3.* 2% of the general population has diabetes. Of these, only 50% are aware of their condition. What is the probability that a randomly selected individual has diabetes and is unaware of it? Let $A$ be the event "person has diabetes" and $B$ be the event "person is unaware of his/her condition". From the statement of the problem we have: $P(A) = 0.02$ and $P(B|A) = 0.50$. Hence

$$P(A \cap B) = P(B|A) \cdot P(A) = (0.50) \cdot (0.02) = 0.01$$

*Example 4.* In the general population of people over 50, about 56% are females. Of these, 12% have arthritis, whereas the percentage of men over 50 who develop this disease is 10%. What is the probability that a randomly selected person over 50 has arthritis? Let $A$ be the event "the person is a female" and $B$ the event "the person has arthritis". We know that $P(A) = 0.56$, $P(B|A) = 0.12$ and $P(B|A') = 0.10$. We have to calculate $P(B)$. Using the Venn diagram (to be drawn in class), we see that

$$P(B) = P(B \cap A) + P(B \cap A')$$

Using the multiplication rule,

$$P(B \cap A) = P(B|A) \cdot P(A) = (0.12) \cdot (0.56) = 0.0672$$

$$P(B \cap A') = P(B|A') \cdot P(A') = (0.10) \cdot (0.44) = 0.044$$

Hence $P(B) = 0.0672 + 0.044 = 0.1112$ (a weighted average between 0.12 and 0.10).

## 4.3. Bayes' Rule

**Bayes' rule** allows us to update the probability of an event (say $B$), in the light of some new information, given by the fact that another event (say $A$) has already happened. The rule says that:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B')P(B')}$$

Note that for the nominator of the fraction above, we used the multiplication rule. For the denominator, we used the following rule, called the **total probability rule**:

$$P(A) = P(A \cap B) + P(A \cap B') = P(A|B)P(B) + P(A|B')P(B')$$

*Example 5.* A screening test is conducted to detect the presence of a medical condition (e.g. a disease). These tests are usually not very accurate. If someone has the disease, the probability that the test will be positive is 80%, whereas if someone does not have the disease, the probability that the test will be negative is 88%. Suppose that 5% of the population has the disease. We would like to calculate the probability that a randomly chosen person will have a positive test.

We denote by $A$ the event "the person has a positive test" and $B$ the event "the person has the disease". We know that $P(A|B) = 0.8$, $P(A|B') = 1 - 0.88 = 0.12$ and $P(B) = 0.05$. Hence

$$P(A) = P(A|B)P(B) + P(A|B')P(B') = (0.8)(0.05) + (0.12)(0.95) = 0.154$$

Using Bayes'rule, we can calculate the probability that a randomly chosen person has the disease, given that this person has a positive test result:

$$P(B|A) = \frac{(0.8)(0.05)}{0.154} = 0.26$$