Math 148 - Intro Stats for Biologists - Fall 2016

Last update: August 28, 2016

Version: **2016-08-28**

Contents

1 Items discussed in lecture but not in the FPP text

1.1 Lists

Given is a list of items $x_1, x_2, x_3, \dots, x_n$. We often write $\vec{x}$ as an abbreviation for such a list [1] , i.e.,

$$\vec{x} = (x_1, x_2, x_3, \dots, x_n) \tag{1.1}$$

It is clear that the last index, here n, denotes for such a list its **size** or **length** There is nothing sacred about the letters x and n. You have seen when we talked about regression that two separate lists $\vec{x} = (x_1, x_2, x_3, \dots, x_n)$ and $\vec{y} = (y_1, y_2, y_3, \dots, y_n)$ of the same length were needed to describe the contents of a scatter diagram. We will also have occasion to look at different lists of different length, so you may see in lecture two separate lists

$$\vec{X} = (X_1, X_2, X_3, \dots, X_n) \quad \text{and} \quad \vec{b} = (b_1, b_2, b_3, \dots, b_k)$$

to describe a chance process $\vec{X}$ of n independent observations $X_1, X_2, \dots, X_n$ each of which will have one of the possible outcomes $b_1, b_2, \dots, b_k$ (the items listed in the "box" [2] $\vec{b}$) with equal probability $1/k$.

Examples for lists:

a. If "H" stands for "Heads" and "T" stands for "Tails" then a possible list that describes the outcomes of 5 flips of a coin would be $\vec{t} = (H, H, T, H, T)$: $t_1 = H, t_2 = H, t_3 = T, t_4 = H, t_5 = T$. We chose "$t$" for "toss" as in toss of a coin. In this example the length of the list $\vec{t}$ is 5.

Lists of such qualitative, non-numeric, variables are very limited in terms of the computations that can be done with them.

[1] That's preferable to writing x because we want to use x only in phrases like "let x be one of the items in the list $\vec{x}$" or as a subscript to an object that is associated with that list (example: SD$_x$, the standard deviation of the list - see further down in this section).

[2] The notation for chance processes and box models will be described more in detail later in this document.

b. Let us rewrite this list as follows: write 1 instead of H and 0 instead of T.

Now we get the list $\vec{x} \ = \ (1,1,0,1,0)$: $x_1 = 1$, $x_2 = 1$, $x_3 = 0$, $x_4 = 1$, $x_5 = 0$.

For such a list we can compute the sum $x_1 + x_2 + \cdots + x_n$ which is $1+1+0+1+0 = 3$ for the list $\vec{x}$ and gives use the number of heads in the original list $\vec{t}$ of heads and tails.

c. If the number of heads is the most important item of interest then we could do the following: create a new list which directly lists for each turn how many heads were obtained until this point in time: We create a new list ("s" for "sum")

$$\begin{aligned}
\vec{s} &= (s_1, s_2, s_3 \dots, s_n) \quad \text{where} \\
s_1 &= x_1, \\
s_2 &= x_1 + x_2, \\
s_3 &= x_1 + x_2 + s_3, \\
&\dots \\
s_n &= x_1 + x_2 + s_3 + \cdots + s_n.
\end{aligned}$$

In the example above which was based on five coin tosses: $s_1 = 1$, $s_2 = 1+1 = 2$, $s_3 = 1+1+0 = 2$, $s_4 = 1+1+0+1 = 3$, $s_5 = 1+1+0+1+0 = 3$.

1.2 "Σ" for Summation

Here is a refresher for the use of "Σ" [3] as a compact means of describing sums of numeric lists.

In section 1.1 on lists we encountered a numbers list

$$\vec{x} \ = \ (1,1,0,1,0): \quad x_1 = 1,\ x_2 = 1,\ x_3 = 0,\ x_4 = 1,\ x_5 = 0$$

and its corresponding list of sums

$$\begin{aligned}
\vec{s} &= (s_1, s_2, s_3, s_4, s_5) \quad \text{where} \\
s_1 &= x_1 \ = \ 1, \\
s_2 &= x_1 + x_2 \ = \ 2, \\
s_3 &= x_1 + x_2 + s_3 \ = \ 2, \\
s_4 &= x_1 + x_2 + s_3 + s_4 \ = \ 3, \\
s_5 &= x_1 + x_2 + s_3 + s_4 + s_5 \ = \ 3.
\end{aligned}$$

Summation occurs so frequently in statistics that we want to have a shorter way to deal with a sum of 3 or 300 or any k numbers $x_1, x_2, \dots, x_k$. We write

$$\sum_{j=1}^{k} x_j \ \text{ as a short for } \ x_1 + x_2 + x_3 + \cdots + x_k \tag{1.2}$$

[3] For additional background check your highschool math books or tutorials on the internet.

We call j the **index variable** and k the **to-variable** or the **end-variable** of the summation. Note that if a variable such as k is used to denote the last item to be included in the sum then it will usually **but not always(!)** be the size of the list.

Let us use this new notation the example above:

$$\begin{aligned}
s_1 &= x_1 = \sum_{j=1}^{1} x_j = 1,\\
s_2 &= x_1 + x_2 = \sum_{j=1}^{2} x_j = 2,\\
s_3 &= x_1 + x_2 + s_3 = \sum_{j=1}^{3} x_j = 2,\\
s_4 &= x_1 + x_2 + s_3 + s_4 = \sum_{j=1}^{4} x_j = 3,\\
.s_5 &= x_1 + x_2 + s_3 + s_4 + s_5 = \sum_{j=1}^{5} x_j = 3.
\end{aligned}$$

Occasionally there are reasons to choose a start index different from 1. For example, we may be interested to sum up the items starting at the 10th turn and ending the 25th turn. We write

$$\sum_{j=10}^{25} x_j \text{ as a short for } x_{10} + x_{11} + x_{12} + \cdots + x_{25}.$$

Sometimes a list may have "start time" 0: $\vec{x} = (x_0, x_1, x_2, \ldots, x_n)$. We then write $\sum_{j=0}^{n} x_j$ for the sum of all its members.

The role of the index variable. In the above the name of the index variable j is irrelevant. You may choose any letter you want as long as it does not match the name of either start variable or end variable. The following all mean the same thing:

$$\sum_{j=4}^{8} x_j = \sum_{i=4}^{8} x_i = \sum_{z=4}^{8} x_z = \sum_{n=4}^{8} x_n = x_4 + x_5 + x_6 + x_7 + x_8$$

and

$$\sum_{j=1}^{n} x_j = \sum_{i=1}^{n} x_i = \sum_{z=1}^{n} x_z,$$

but it is **illegal** to write $\sum_{n=1}^{n} x_n$ because you may not mix up the index variable n with the end variable n in this summation expression.

Lazy ways to write the Σ notation: If there is absolutely no confusion about start index and end index (practically because start index = first index of the list (usually 1) and end index = last index of the list) then these may be dropped and the following each mean the same:

(1.3) For a list $\vec{x} = (x_a, x_{a+1}, x_{a+2}, \dots x_{n-1}, x_{n-1})$, $\displaystyle \sum x_j \ := \ \sum_j x_j \ := \ \sum_{j=a}^{n} x_j.$

2 Math 148 - Various topics

Various notes on all kind of subjects.

2.1 Function notation

Remark 2.1 (Functions: from the concrete to the abstract)**.**

a. $y = f(x) = 3x^2 - 5$: assigns the argument or independent variable x to its function value $f(x)$ "$\mapsto$" means "assigns to", so in short: $x \mapsto 3x^2 - 5$, also $x \mapsto f(x)$

b. x and y are just dummy variables. It does not mean what they are called. We could have written $x = f(t) = 3t^2 - 5$ instead! Short form: $t \mapsto 3t^2 - 5$, also $t \mapsto f(t)$

c. Because we do not deal with some function in the abstract but the concrete one which does the assignment $x \mapsto 3x^2 - 5$, it does not matter whether we write $y = f(x) = 3x^2 - 5$ or $y = H(x) = 3x^2 - 5$ or $v = H(u) = 3u^2 - 5$. It's still one and the same function which squares its argument, multiplies that by 3 and then subtracts 5.

d. Functions need not necessarily have numbers for arguments and or function values but we might have "vectors" (finite lists of numbers) instead: The function

$$(x_1, x_2, x_3) \mapsto F(x_1, x_2, x_3) := \Big(\ (x_1 + x_2 + x_3)/3,\ \sqrt{x_1^2 + x_2^2 + x_3^2}\ \Big)$$

throws any triplet of numbers (a 3–dimensional vector) into a pair of numbers (a 2–dimensional vector).

For example, $F(1, -2, 3) = \Big(\ (1 - 2 + 3)/3,\ \sqrt{1^2 + (-2)^2 + 3^2}\ \Big) = (2/3, \sqrt{14})$.

e. Now comes a big jump in abstraction: Sometimes we neither know nor care about the nature of the arguments x and the pool or source from which they are drawn. In math we do not talk about "pools". Instead we talk about sets. A **set** is for simply a collection of stuff. Don't try to overcomplicate that. For example, we talk about the set of all numbers or that of all 3–dimensional vectors or the set $]3.8, 17[$ of all numbers between 3.8 and 17, endpoints excluded or the set $[3.8, 17]$ of all numbers between 3.8 and 17, endpoints included.

f. Particularly popular for the set of all arguments: Ω (uppercase Omega). Particularly popular for the arguments instead: ω instead of x or t. So now we talk about a function $\omega \mapsto y := f(\omega)$ with arguments being taken from Ω.

g. Rather than using "f" for the function name and y for the name of the function value you will often see "X" or "Y" or "S" for the function name and x or y or s for the name of the function value. In other words we often deal with

$$\omega \mapsto x = X(\omega), \quad \omega \mapsto y = Y(\omega), \quad \omega \mapsto s = S(\omega), \ldots$$

.

h. What will that be good for? We will think of X something which describes an outcome x which depends on a randomly happening phenomenon ω. For example X might mean rolling a die. The outcome x as the result of rolling the die will be one of the numbers $1, 2, 3, 4, 5, 6$. We don't know beforehand which one it will be because that's happening by chance which is represented by ω.

Think of it this way if you like: Some supreme being picks some ω from the set Ω. If it happens to be "this" ω_1 then the outcome will be $x_1 = X(\omega_1)$ which happens to be 5. If "that" ω_2 had been picked then the outcome will be $x_2 = X(\omega_2)$. which happens to be 1 instead. We often don't know and don't need to know for the questions that interest us anything more specific about the nature of Ω and X, and how X assigns ω to some value between 1 and 6.

2.2 FPP ch.1,2: experimental design and confounding

Definition 2.1 (Design of Experiments)**.** [4]

Design of experiments, also referred to as **experimental Design** is the planning process that needs to go into any experiment, survey or study in order for the results to be valid. The medical and social sciences tend to use the term "Experimental Design" while engineering, industrial and computer sciences favor the term "Design of experiments."

Design of experiments involves:

a. The systematic collection of data
b. A focus on the design itself, rather than the results
c. Planning changes to independent (input) variables and the effect on dependent variables or response variables
d. Ensuring results are valid, easily interpreted, and definitive.

The most important principles are:

a. **Randomization**: the selection of data by a completely random method, like simple random sampling. Randomization significantly eliminates bias.
b. **Replication**: the experiment must be repeatable by other researchers.

Design of Experiments: Categories

a. **Cross Sectional Study**:
b. **Longitudinal study**:
c. **Observational study**:

Variables in Design of Experiments

[4] Source: http://www.statisticshowto.com/design-of-experiments/.

a. Confounding variables
b. Control variables
c. Dependent variables
d. Explanatory variables
e. Outcome variables

Remark 2.2 (Advantages and disadvantages of randomized Controlled experiments). [5]

Advantages

a. Random allocation can cancel out population bias; it ensures that any other possible causes for the experimental results are split equally between groups.
b. Blinding is easy to include in this type of experiment.
c. Results from the experiment can be analyzed with statistical tests and used to infer their validity for the entire population.

Disadvantages

a. Generally more expensive and more time consuming than other methods.
b. Very large sample sizes (over 5,000 participants) are often needed.
c. Random controlled trials cannot uncover causation/risk factors. For example, ethical concerns would prevent a randomized controlled trial investigating the risk factors for smoking.
d. Some programs, for example cancer screening, are unsuited for random allocation of participants (again, due to ethical concerns).
e. Volunteer bias can be an issue.

Example 2.1 (Randomized controlled experiment). [6]

To determine how a new type of short wave UVA-blocking sunscreen affects the general health of skin in comparison to a regular long wave UVA-blocking sunscreen, 40 trial participants were randomly separated into equal groups of 20: an experimental group and a control group. All participants' skin health was then initially evaluated. The experimental group wore the short wave UVA-blocking sunscreen daily, and the control group wore the long wave UVA-blocking sunscreen daily.

After one year, the general health of the skin was measured in both groups and statistically analyzed. In the control group, wearing long wave UVA-blocking sunscreen daily led to improvements in general skin health for 60% of the participants. In the experimental group, wearing short wave UVA-blocking sunscreen daily led to improvements in general skin health for 75% of the participants.

Some questions to be answered later in this course:

[5] Source: http://www.statisticshowto.com/experimental-design/#RandomC.
[6] Source: https://himmelfarb.gwu.edu/tutorials/studydesign101/rcts.html.

a. Is the increase in improvement rates from 60% to 75% significant enough to make it practically unlikely that it resulted from the "luck of the draw" as far as the selection of the participants and/or their allocation to treatment or control was concerned?
b. Can we infer the answers to question **a** from the information given or do we need to know more about the data?

Definition 2.2 (Confounding variables)**.** [7]

In an experiment, the independent variable typically has an effect on your dependent variable. For example, if you are researching whether lack of exercise leads to weight gain, lack of exercise is your independent variable and weight gain is your dependent variable. A **confounding variable** is any other variable that also has an effect on your dependent variable. Confounding variables are like extra independent variables that are having a hidden effect on your dependent variables.

Another way to express this is that confounding variables cause the so-called **third variable problem** which refers to the fact that any time we observe a relationship among two variables, there's always the possibility that some third variable which we don't know about is responsible for ("confounding") the relationship.

Confounding variables can cause major problems:

a. They introduce bias.
b. They damage the validity of the results gained in the experiment because they indicate a relationship between two variables which does not really exist.

Example 2.2 (Confounding variable)**.** [8]

You test 200 volunteers (100 men and 100 women). You find that lack of exercise leads to weight gain. Could there be problems with your design?

One problem with your experiment is that is lacks any control variables. For example, the use of placebos, or random assignment to groups. So you really can't say for sure whether lack of exercise leads to weight gain. One confounding variable is how much people eat. It's also possible that men eat more than women; this could also make sex a confounding variable. Nothing was mentioned about starting weight, occupation or age either. A poor study design like this could lead to bias. For example, if all of the women in the study were middle-aged, and all of the men were aged 16, age would have a direct effect on weight gain. That makes age a confounding variable.

Example 2.3 (Confounding variable)**.** [9]

A research group designs a a study to determine if heavy drinkers die at a younger age.

They gather the data. Their results, and a battery of statistical tests, indeed show that people who drink excessively are likely to die younger. Unfortunately, when the researchers do a crosscheck with their peers, the results are ripped apart, because their peers live just as long. Could there be another factor, not measured, that influences both drinking and living age?

[7] Source: http://www.statisticshowto.com/design-of-experiments/confounding-variable/. This link contains a very good VIDEO!

[8] Source: http://www.statisticshowto.com/design-of-experiments/confounding-variable/.

[9] Source: https://explorable.com/confounding-variables.

The weakness in the experimental design was that it did not take into account the confounding variables, and did not try to eliminate or control any other factors. For example, it is quite possible that the heaviest drinkers hailed from a different background or social group. Heavy drinkers may be more likely to smoke, or eat junk food, all of which could be factors in reducing longevity.

Example 2.4 (Confounding variable)**.** [10]

Michael conducts an experiment to test the effectiveness of a pain reliever. He gives the pain reliever to ten people in the experiment and herbal tea to another ten people. Unfortunately, all of the people in the study, the ten that took the pain reliever and the ten that took the herbal tea, report improvements in their headaches.

The results could be due to a confounding variable: Michael does not have a control group. If he had a group that took nothing for the headaches, he could make a more accurate analysis. It is possible that the herbal tea has healing properties for pain. This is a confounding variable which messes up the reliability of the results as far as answering the question whether or not the pain relievers are effective is concerned.

Example 2.5 (Confounding variables - MURDER AND ICE CREAM)**.** [11]

It is known that throughout the year, murder rates and ice cream sales are highly positively correlated. That is, as murder rates rise, so does the sale of ice cream. There are three possible explanations for this kind of relationship:

a. Murders cause people to purchase ice cream. Perhaps when one is murdered, they are resurrected as zombies who primarily feed on ice cream.

b. Purchasing ice cream causes people to murder or get murdered. Perhaps when one eats ice cream, those without ice cream become jealous and murder those with ice cream.

c. There is a confounding variable which causes the increase in BOTH ice cream sales AND murder rates. For instance, the weather. When it's cold and wintery, people stay at home rather than go outside to murder people. They also probably don't eat a lot of ice cream. On the other hand, when it's hot and summery, people spend more time outside interacting with each other, and hence are more likely to get into the kinds of situations that lead to murder. They are also probably buying ice cream.

In the above example, the weather is a variable that confounds the relationship between ice cream sales and murder rates.

A CONCRETE EXAMPLE WITH A PRETTY PICTURE

A sadly confounded experimental design

If you still don't get it, I've got another example and a pretty picture to go with it. Imagine you'd like to examine the relationship between the force you apply to a ball and the distance the ball travels. Naturally, you predict that the more force you apply, the further the ball will travel. You create an experiment with two conditions. Now you should be eyeballing that pretty picture I told you about. In Condition 1, you apply a 10 lb force to the ball and measure the distance traveled. In Condition 2, you apply a 5 lb force to the ball and measure the distance traveled. After you run your

[10] Source: https://explorable.com/confounding-variables.

[11] Source: http://www.psychologyinaction.org/2011/10/30/what-is-a-confounding-variable/.

experiment, you observe that the ball travels further in Condition 2 than it does in Condition 1. In other words, you find that the less force you apply, the further the ball travels. Should you conclude that Isaac Newton was wrong? No. As should be pretty clear from that pretty picture, there's a clear confounding variable in this experimental design: the angle of the slope. Given the presence of this confound, we have no way of knowing which variable – force or angle – is responsible for the change in the distance the ball travels.

This illustrates why it's so important to always be on the lookout for confounds. Confounds can make us reach conclusions that are wrong; confounds can make us look stupid. Or, to use slightly more technical language: confounding variable = BAD.

Remark 2.3 (Correlation and causation)**.** [12]

Confounding variables are closely related to the problem of correlation and causation.

For example, a scientist performs statistical tests, sees a correlation and incorrectly announces that there is a causal link between two variables. Constant monitoring, before, during and after an experiment, is the only way to ensure that any confounding variables are eliminated.

[12] Source: https://explorable.com/confounding-variables.

References

List of Symbols

$\complement A$ - complement, 2
$\sum_{j=1}^{k} x_j$ - summation, 4
$\vec{x}$ - list of items, 2

Index