Introduction to Statistics and Experimental Design

Reuben Thomas
bioinformatics@gladstone.ucsf.edu

February 17, 2021

We are all in the business of identifying/determining new scientific phenomena or proposing mechanisms for given processes. We generate empirical data from experiments to support our conclusions. The conclusions from these data are meant to support two target audiences:

1. Ourselves, our PI and lab.
2. The broader scientific community who will read about our work in scientific journals

The data we obtain from our experiments are almost always random but hopefully not arbitrary, so measurement of a particular gene's expression will be different from sample to sample, even within the sample there may be variation depending on the time of the day, the month the experiment is performed, the person doing the experiment, the reagents used. Despite these variations associated with the measured gene expression we typically observe that there is something common in these measurements. For example, the mean measurement across different sample may be more or less irrespective for whether one use these 10 animals or another set of 10 animals.

Suppose we would like to claim that a particular gene's expression is elevated in a particular tissue with a drug treatment in mice. We may base this claim based on data from one randomly sampled mouse treated with the drug and from another mouse treated with the placebo. Unfortunately, given all the above reasons why the level measured once in once mouse could vary, this claim would be viewed with great skepticism by most reviewers. The only situation where these data may be useful would be these two were really special one-of-a-kind mice who you would probably name like ice-preserved cave men and women from prehistoric times. In these situations, the claim would be specific to these special mice or cave men.

If you would like to make a claim that is applicable across the whole population of mice (what is generally implicit in your claim) then the only way to convince the skeptic is to generate data across a group of randomly sampled mice. How many you need to sample really depends on the underlying variability of this gene expression across the population, across the different times it is measured etc. This is where the magic of statistics comes in - with a limited subset

of samples you will be able to make claims about the whole population.

There are three main concepts that you (or people reviewing your claim) need to considers when designing your experiments to generate the data necessary to support your claim: **CHANCE**, *CONFOUNDING* and **BIAS**.

The use of more number of samples in a given (treatment) condition would help in convincing someone that the claim you have made is not the result of **CHANCE**. The two other concepts are also very important. You don't want to be in a situation when all mice given the treatment are not male and mice not given treatment are female, all samples from one condition processed in January while all samples from the other condition processed in February. You want to make sure that the treatment is not confounded with another variable that affects gene expression. Issues of bias could arise when we are near the edges of the dynamic range of the assay we use. So if want to claim a gene's expression is not affected by the treatment we need to make sure the level of this gene's expression is within the dynamic range of the assay. Your claims may be doubted if you haven't adequately addressed these issues of **BIAS** and **CONFOUNDING** even if you have 100 replicates in each condition.

Practical tips in designing your own experiment so that your conclusions are reproducible and correctly interpreted

You are interested in testing the association between a factor (e.g., genotype, drug treatment, disease) and a given response (e.g., gene expression, behaviour).

1. Create a list of biological factors (that you are not really interested in) that could affect the response (e.g, gender, bmi)

2. Create a litst of non-biological or technical factors (that are definitely not interested in) that could affect the response (e.g., time/day/month of experiment/batch, reagents used, reagents batch used, person doing the experiment, lane on the sequencer)

3. What is your target population on which you would like to base your conclusions? Decide on the generality of the conclusions that you would like to draw - do you want to suggest that your hypothesis is valid across all mice of a given strain? across all subjects diagnosed with autism spectrum disorder? across all subjects diagnosed with autism spectrum disorder in a particular geographical area? across all subjects diagnosed with autism spectrum disorder below the 5 years of age?

4. Look at the pool or population of samples you have available. Do they meet the generality of the conclusions you desire? If not, then under what assumptions can we make the general conclusions.

5. Decide on the number of samples (to be drawn from the population of samples) under each setting. You can either base it on similar experiments performed in the past (either in your lab, institute or literature) or get advice from a statistician. In either case, this number should be almost always at least 3 - the minimum number necessary to get a reasonable estimate of the variability of the response. The statistician could suggest a larger number of replicates if you believe that the effect of the factor of interest on the response is relatively small.

6. *Randomization*: is a key principle that will help you generate good experimental designs that avoids problems of confounding due to biological and technical factors. When you pick samples from your target population, make sure that they are randomly picked - you don't want to pick the "larger/redder/healthier-looking" samples. When you assign treatment to a sample, make sure you randomly assign treatment to the samples. So you don't want to assign the drug treatment to all animals sampled today and the control/DMSO treatment to all animals sampled the next week. If all samples cannot be proceesed on one day then needs to processed in two batches (or say lanes on a sequencer) then make sure that each batch has samples from each condition after you have randomly assigned samples to batches. With smaller number of samples in each condition it is sometimes better go for "balanced" designs. So if you want your conclusions to be applicable to the entire population of animals regardless of gender then you could randomly sample half the replicate number of animals from a given gender in each condition.

7. Once you have performed your experiment and collected the data then you will need to decide on a statistic and kind of hypothesis test to perform. Here again use the past knowledge in your lab/literature or consult a statistician.

Examples

As we go over these examples that all involve claims/conclusions drawn from different statistical tests, please try and answer the following questions. Note, some of these claims are meant to capture some of the typical headlines one sees in the news, headlines that would not stand scrutiny if one carefully looks at the data.

1. What is the target population? Or how generally applicable are the conclusions? Or would (the sceptical) you buy this claim?
2. What is the null hypothesis or the sceptical point of view?
3. Is the p-value reported reflective of the way the subjects were sampled?
4. Can you think of some potential confounders that could affect the conclusions?

Census data

Claim: A random person you meet in country X is more likely to be an atheist than a person you meet in country Y.

Data: Census data from each of these countries. Information from 100% of the populations from the two countries were recorded. The ratio of the odds of a person in the country X being an atheist versus the odds of a person from country Y being an atheist is 20. The associated p-value is 10^{-6}.

Autism spectrum disorder data

Claim: Potential diagnostic for ASD? Marker X is elevated in subjects diagnosed with ASD compared with subjects not diagnosed with ASD.

Data: Subjects below 5 years were drawn from patients (coming to UCSF) on a volunteer basis. The associated p-value is 0.1.

Cell survival in disease

Claim: Interneuronal cells in patients with disease X have lower survival compared to those in patients without the disease.

Data: Skin cells from patient John Doe were reprogrammed to interneuronal cells. Skin cells from healthy subject Mark Thomas of the same age and gender as John Doe were also reprogrammed to interneuronal cells. Interneuronal cells from these two subjects were placed in two 96-well plates. Cell survival times were noted for cells in each of the wells. The hazard ratio of cell death at any time is 3.0 for cells belonging to John Doe versus those belonging to Mark Thomas. The associated p-value is 0.02.

100 differentially regulated immune functon-related genes due to environmental exposure to benzene

Claim: Exposure to benzene leads to altered immune function.

Data: The gene expression levels in the Peripheral Blood Mononuclear Cells (PBMCs) in 10 subjects occupationally exposed to benzene in a shoe-manufacturing plant in Tianjin, China and 10 subjects not likely exposed to benzene in another factory in the area were assayed using Illumina BeadChip microarrrays. Of the 3000 genes identified as differentially expressed using a multiple testing adjusted p-value < 0.05 threshold, 100 genes are annotated as having some immune function.

Heterogenity from cellular reprogramming

Claim: Reprogramming cardiac fibroblasts to cardiomyocytes following Protocol A results in at least 6 different cell-types at day 17 that include cardiomyocyte-like cells.

Data: Neonatal cardiac fibroblasts from five different mouse embryos were pooled together and subjected to Protocol A. 10,000 cells at day 17 were run on the 10X genomics scRNA-seq platform. The clustering of the cells by their gene expression profiles reveals 6 distinct clusters to cells.

Altered gene expression in a subset of cardiac cells due to mutation in a developmental gene, X

Claim: 5% of cardiomyocytes have altered gene expression due to mutation in gene X.

Data: Gene expression profiles of 2000 cardiomyocytes each taken from a wild-type mouse and from a mouse having a mutatioin gene X are assayed using scRNA-seq. The combined 4000 cells are clustered revealing one cluster of 200 cells made up solely of mutant cells. After correcting for multiple testing (p-value < 0.05), 257 genes are identified that are differentially expressed between this group of cells versus the rest of the cells.

Pathway/Gene Ontology enrichment analyses

Claim: Altered immune function is associated with benzene exposure in humans.

Data: is the same the benzene example above. The list of 3000 differentially expressed genes is submitted to a Gene Ontology enrichment program DAVID. The top enriched GO process is the immune function GO biological process.