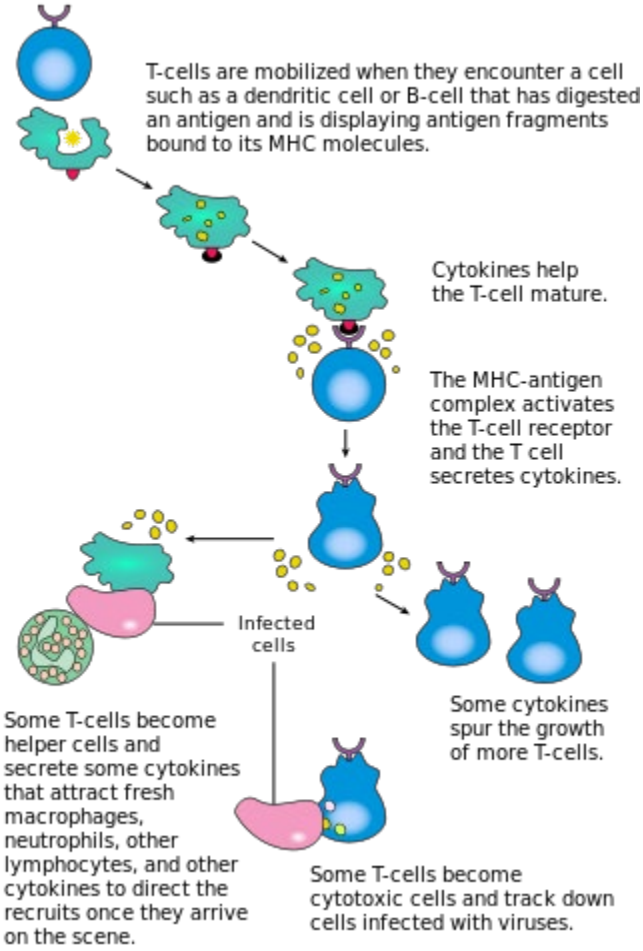
Genetic and Molecular Epidemiology Notes

Immune System Quick Review

The MHC region contains numerous polymorphic and multicopy genes that play important roles not only in tissue histocompatibility but also other primary functions in the immune system that provide protection against pathogens. The first MHC products were discovered on the surface of leucocytes (white blood cells) and the human MHC was initially referred to as the human leucocyte antigen (HLA) complex. Human MHC genes are categorized into three classes: MHC classes I, II and III. MHC class I (HLA-A, -B and -C) and class II (HLA-DR, -DQ and -DP) genes encode antigen-presenting molecules that are expressed on antigen-presenting cells and stimulate CD8+ and CD4+ T cells, respectively. MHC class III or central MHC proteins are a diverse group of molecules that perform various immune functions in the body such as complement proteins involved in the antibody response, and inflammatory cytokines.

Q: What is the difference between NK and CTL if CTL do not need additional antigen presentation or recognition?



T Cells

A T cell, or T lymphocyte, is a type of lymphocyte (a subtype of white blood cell) that plays a central role in cell-mediated immunity. T cells can be distinguished from other lymphocytes, such as B cells and natural killer cells, by the presence of a T-cell receptor on the cell surface. They are called *T cells* because they mature in the thymus from thymocytes[1] (although some also mature in the tonsils). The several subsets of T cells each have a distinct function. The majority of human T cells rearrange their alpha and beta chains on the cell receptor and are termed alpha beta T cells (αβ T cells) and are part of the adaptive immune system. Specialized gamma delta T cells, (a small minority of T cells in the human body, more frequent in ruminants), have invariant T-cell receptors with limited diversity, that can effectively present antigens to other T cells and are considered to be part of the innate immune system.

Cytotoxic T Cells

Cytotoxic T cells (T_C cells, CTLs, T-killer cells, killer T cells) destroy virus-infected cells and tumor cells, and are also implicated in transplant rejection. These cells are also known as $CD8^+$ T cells since they express the CD8 glycoprotein at their surfaces. These cells recognize their targets by binding to antigen associated with MHC class I molecules, which are present on the surface of all nucleated cells. Through IL-10, adenosine, and other molecules secreted by regulatory T cells, the $CD8^+$ cells can be inactivated to an anergic state, which prevents autoimmune diseases.

Most cytotoxic T cells express T-cell receptors (TCRs) that can recognize a specific antigen. An antigen is a molecule capable of stimulating an immune response, and is often produced by cancer cells or viruses. Antigens inside a cell are bound to class I MHC molecules, and brought to the surface of the cell by the class I MHC molecule, where they can be recognized by the T cell. If the TCR is specific for that antigen, it binds to the complex of the class I MHC molecule and the antigen, and the T cell destroys the cell.

In order for the TCR to bind to the class I MHC molecule, the former must be accompanied by a glycoprotein called CD8, which binds to the constant portion of the class I MHC molecule. Therefore, these T cells are called CD8+ T cells.

The affinity between CD8 and the MHC molecule keeps the T_C cell and the target cell bound closely together during antigen-specific activation. CD8+ T cells are recognized as T_C cells once they become activated and are generally classified as having a pre-defined cytotoxic role within the immune system. However, CD8+ T cells also have the ability to make some cytokines.

The vast majority of T cells express alpha-beta TCRs (αβ T cells), but some T cells in epithelial tissues (like the gut) express gamma-delta TCRs (γδ T cells), which recognize non-protein antigens.

T cells with functionally stable TCRs express both the CD4 and CD8 co-receptors and are therefore termed "double-positive" (DP) T cells (CD4+CD8+). The double-positive T cells are exposed to a wide variety of self-antigens in the thymus and undergo two selection criteria:

1. positive selection, in which those double-positive T cells that bind to foreign antigen in the presence of self MHC. They will differentiate into either CD4+ or CD8+ depending on which MHC is associated with the antigen presented (MHC1 for CD8, MHC2 for

CD4). In this case, the cells would have been presented antigen in the context of MHC1. Positive selection means selecting those TCRs capable of recognizing self MHC molecules.

2. negative selection, in which those double-positive T cells that bind *too strongly* to MHC-presented *self antigens* undergo apoptosis because they could otherwise become autoreactive, leading to autoimmunity.

Only those T cells that bind to the MHC-self-antigen complexes weakly are positively selected. Those cells that survive positive and negative selection differentiate into single-positive T cells (either CD4+ or CD8+), depending on whether their TCR recognizes an MHC class I-presented antigen (CD8) or an MHC class II-presented antigen (CD4). It is the CD8+ T-cells that will mature and go on to become cytotoxic T cells following their activation with a class I-restricted antigen.

With an exception of some cell types, such as non-nucleated cells (including erythrocytes), Class I MHC is expressed by all host cells. When these cells are infected with a virus (or another intracellular pathogen), the cells degrade foreign proteins via antigen processing. These result in peptide fragments, some of which are presented by MHC Class I to the T cell antigen receptor (TCR) on CD8+ T cells.

The activation of cytotoxic T cells is dependent on several simultaneous interactions between molecules expressed on the surface of the T cell and molecules on the surface of the antigen-presenting cell (APC).

A simple activation of naive $CD8^+$ T cells requires the interaction with professional antigen-presenting cells, mainly with matured dendritic cells. To generate long lasting memory T cells and to allow repetitive stimulation of cytotoxic T cells, dendritic cells have to interact with both, activated $CD4^+$ helper T cells and $CD8^+$ T cells. During this process, the $CD4^+$ helper T cells "license" the dendritic cells to give a potent activating signal to the naive $CD8^+$ T cells.

Once activated, the T_C cell undergoes clonal expansion with the help of the cytokine Interleukin-2 (IL-2), which is a growth and differentiation factor for T cells. This increases the number of cells specific for the target antigen that can then travel throughout the body in search of antigen-positive somatic cells.

Dendritic cell

Dendritic cells (DCs) are antigen-presenting cells (also known as *accessory cells*) of the mammalian immune system. Their main function is to process antigen material and present it on the cell surface to the T cells of the immune system. They act as messengers between the innate and the adaptive immune systems. Dendritic cells are present in those tissues that are in contact with the external environment, such as the skin (where there is a specialized dendritic cell type called the Langerhans cell) and the inner lining of the nose, lungs, stomach and intestines. They can also be found in an immature state in the blood. Once activated, they migrate to the lymph nodes where they interact with T cells and B cells to initiate and shape the adaptive immune response.

The dendritic cells are constantly in communication with other cells in the body. This communication can take the form of direct cell–cell contact based on the interaction of cell-surface proteins, or at a distance via cytokines. For example, stimulating dendritic cells *in vivo* with microbial extracts causes the dendritic cells to rapidly begin producing IL-12. IL-12 is

a signal that helps send naive CD4 T cells towards a Th1 phenotype. The ultimate consequence is priming and activation of the immune system for attack against the antigens which the dendritic cell presents on its surface.

Helper T cells

T helper cells (T_H cells) assist other white blood cells in immunologic processes, including maturation of B cells into plasma cellsand memory B cells, and activation of cytotoxic T cells and macrophages. These cells are also known as $CD4^+$ T cells because they express the CD4 glycoprotein on their surfaces. Helper T cells become activated when they are presented with peptide antigens by MHC class II molecules, which are expressed on the surface of antigen-presenting cells (APCs). Once activated, they divide rapidly and secrete small proteins called cytokines that regulate or assist in the active immune response. These cells can differentiate into one of several subtypes, including T_H1, T_H2, T_H3, T_H17, T_H9, or T_{FH}, which secrete different cytokines to facilitate different types of immune responses. Signalling from the APC directs T cells into particular subtypes.

Memory T cells

Antigen-naïve T cells expand and differentiate into memory and effector T cells after they encounter their cognate antigen within the context of an MHC molecule on the surface of a professional antigen presenting cell (e.g. a dendritic cell). Appropriate co-stimulation must be present at the time of antigen encounter for this process to occur. Historically, memory T cells were thought to belong to either the effector or central memory subtypes, each with their own distinguishing set of cell surface markers (see below).[5] Subsequently, numerous new populations of memory T cells were discovered including tissue-resident memory T (Trm) cells, stem memory TSCM cells, and virtual memory T cells. The single unifying theme for all memory T cell subtypes is that they are long-lived and can quickly expand to large numbers of effector T cells upon re-exposure to their cognate antigen. By this mechanism they provide the immune system with "memory" against previously encountered pathogens. Memory T cells may be either $CD4^+$ or $CD8^+$ and usually express CD45RO.

Regulatory (suppressor)

Regulatory T cells (suppressor T cells) are crucial for the maintenance of immunological tolerance. Their major role is to shut down T cell-mediated immunity toward the end of an immune reaction and to suppress autoreactive T cells that escaped the process of negative selection in the thymus. Suppressor T cells along with Helper T cells can collectively be called Regulatory T cells due to their regulatory functions. Two major classes of $CD4^+$ T_{reg} cells have been described — $FOXP3^+$ T_{reg} cells and $FOXP3^-$ T_{reg} cells.

Regulatory T cells can develop either during normal development in the thymus, and are then known as thymic Treg cells, or can be induced peripherally and are called peripherally derived Treg cells. These two subsets were previously called "naturally occurring", and "adaptive" or "induced", respectively.

Gamma delta T cells

Gamma delta T cells (γδ T cells) represent a small subset of T cells that possess a distinct T cell receptor (TCR) on their surfaces. A majority of T cells have a TCR composed of

two glycoprotein chains called α- and β- TCR chains. However, in γδ T cells, the TCR is made up of one γ-chain and one δ-chain. This group of T cells is much less common in humans and mice (about 2% of total T cells); and are found mostly in the gut mucosa, within a population of lymphocytes known as intraepithelial lymphocytes. In rabbits, sheep, and chickens, the number of γδ T cells can be as high as 60% of total T cells. The antigenic molecules that activate γδ T cells are still widely unknown. However, γδ T cells are not MHC-restricted and seem to be able to recognize whole proteins rather than requiring peptides to be presented by MHC molecules on APCs.

Natural killer T cell

Natural killer T cells (NKT cells – not to be confused with natural killer cells of the innate immune system) bridge the adaptive immune system with the innate immune system. Unlike conventional T cells that recognize peptide antigens presented by major histocompatibility complex (MHC) molecules, NKT cells recognize glycolipid antigen presented by a molecule called CD1d. Once activated, these cells can perform functions ascribed to both T_h and T_c cells (i.e., cytokine production and release of cytolytic/cell killing molecules). They are also able to recognize and eliminate some tumor cells and cells infected with herpes viruses

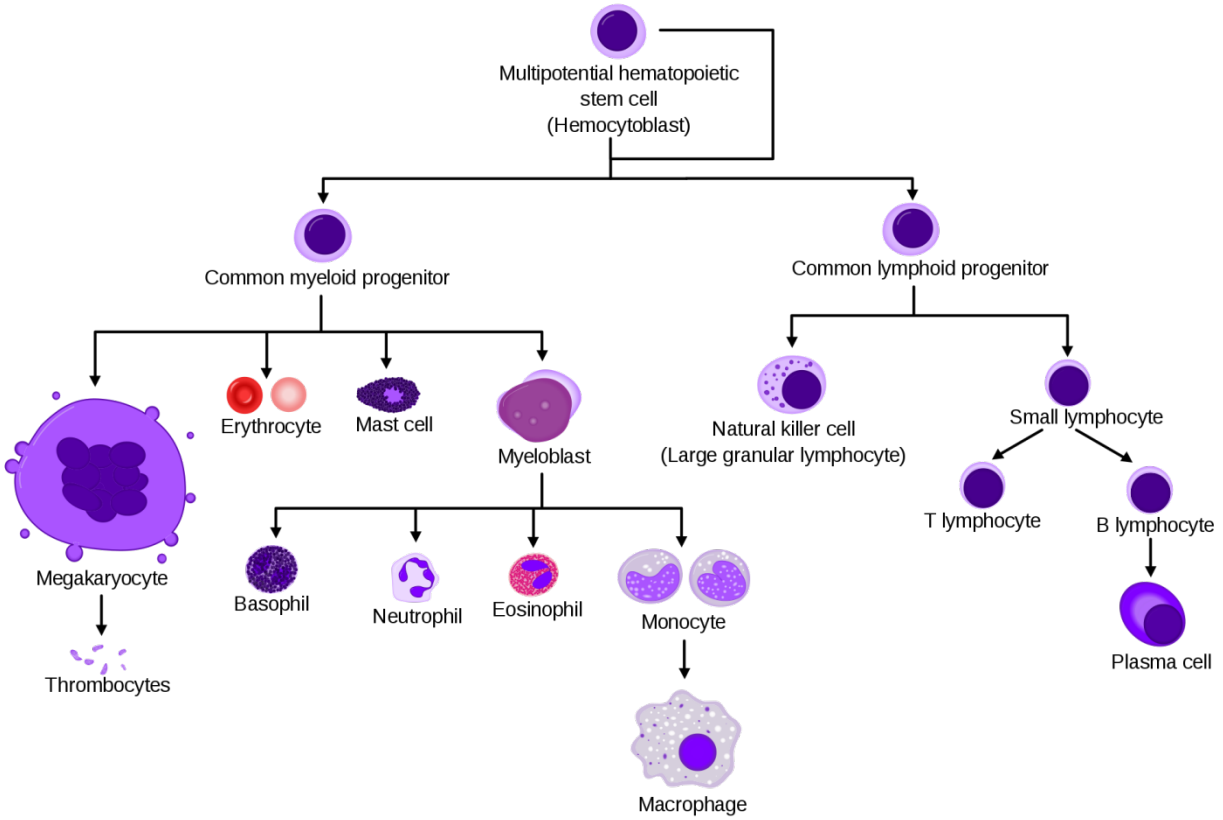
Natural Killer Cells

Natural killer cells or NK cells are a type of cytotoxic lymphocyte critical to the innate immune system. The role NK cells play is analogous to that of cytotoxic T cells in the vertebrate adaptive immune response. NK cells provide rapid responses to viral-infected cells, acting at around 3 days after infection, and respond to tumor formation. Typically, immune cells detect major histocompatibility complex (MHC) presented on infected cell surfaces, triggering cytokine release, causing lysis or apoptosis. NK cells are unique, however, as they have the ability to recognize stressed cells in the absence of antibodies and MHC, allowing for a much faster immune reaction. They were named "natural killers" because of the initial notion that they do not require activation to kill cells that are missing "self" markers of MHC class 1. This role is especially important because harmful cells that are missing MHC I markers cannot be detected and destroyed by other immune cells, such as T lymphocyte cells.

NK cells (belonging to the group of innate lymphoid cells) are defined as large granular lymphocytes (LGL) and constitute the third kind of cells differentiated from the common lymphoid progenitor-generating B and T lymphocytes. NK cells are known to differentiate and mature in the bone marrow, lymph nodes, spleen, tonsils, and thymus, where they then enter into the circulation. NK cells differ from natural killer T cells (NKTs) phenotypically, by origin and by respective effector functions; often, NKT cell activity promotes NK cell activity by secreting interferon gamma. In contrast to NKT cells, NK cells do not express T-cell antigen receptors (TCR) or pan T marker CD3 or surface immunoglobulins (Ig) B cell receptors.

In addition to the knowledge that natural killer cells are effectors of innate immunity, recent research has uncovered information on both activating and inhibitory NK cell receptors which play important functional roles, including self-tolerance and the sustaining of NK cell activity. NK cells also play a role in the adaptive immune response:[6] numerous experiments have demonstrated their ability to readily adjust to the immediate environment and formulate antigen-specific immunological memory, fundamental for responding to secondary infections with the

same antigen.[7] The role of NK cells in both the innate and adaptive immune responses is becoming increasingly important in research using NK cell activity as a potential cancer therapy.



B cells

B cells, also known as B lymphocytes, are a type of white blood cell of the lymphocyte subtype. They function in the humoral immunity component of the adaptive immune system by secreting antibodies. Additionally, B cells present antigen (they are also classified as professional antigen-presenting cells (APCs)) and secrete cytokines.

B cells, unlike the other two classes of lymphocytes, T cells and natural killer cells, express B cell receptors (BCRs) on their cell membrane. BCRs allow the B cell to bind a specific antigen, against which it will initiate an antibody response

- Plasmablast - A short-lived, proliferating antibody-secreting cell arising from B cell differentiation. Plasmablasts are generated early in an infection and their antibodies tend to have a weaker affinity towards their target antigen compared to plasma cell. Plasmablasts can result from T cell-independent activation of B cells or the extrafollicular response from T cell-dependent activation of B cells.
- Plasma cell - A long-lived, non-proliferating antibody-secreting cell arising from B cell differentiation. There is evidence that B cells first differentiate into a plasmablast-like cell, then differentiate into a plasma cell. Plasma cells are generated later in an infection and, compared to plasmablasts, have antibodies with a higher affinity towards their target

antigen due to affinity maturation in the germinal center (GC) and produce more antibodies. Plasma cells typically result from the germinal center reaction from T cell-dependent activation of B cells, however they can also result from T cell-independent activation of B cells.

- Lymphoplasmacytoid cell - A cell with a mixture of B lymphocyte and plasma cell morphological features that is thought to be closely related to or a subtype of plasma cells. This cell type is found in pre-malignant and malignant plasma cell dyscrasias that are associated with the secretion of IgM monoclonal proteins; these dyscrasias include IgM monoclonal gammopathy of undetermined significance and Waldenström's macroglobulinemia.
- Memory B cell - Dormant B cell arising from B cell differentiation. Their function is to circulate through the body and initiate a stronger, more rapid antibody response (known as the secondary antibody response) if they detect the antigen that had activated their parent B cell (memory B cells and their parent B cells share the same BCR, thus they detect the same antigen). Memory B cells can be generated from T cell-dependent activation through both the extrafollicular response and the germinal center reaction as well as from T cell-independent activation of B1 cells.

T cell-dependent activation

Antigens that activate B cells with the help of T-cell are known as T cell-dependent (TD) antigens and include foreign proteins. They are named as such because they are unable to induce a humoral response in organisms that lack T cells. B cell response to these antigens takes multiple days, though antibodies generated have a higher affinity and are more functionally versatile than those generated from T cell-independent activation.

Once a BCR binds a TD antigen, the antigen is taken up into the B cell through receptor-mediated endocytosis, degraded, and presented to T cells as peptide pieces in complex with MHC-II molecules on the cell membrane. T helper (TH) cells, typically follicular T helper (TFH) cells, that were activated with the same antigen recognize and bind these MHC-II-peptide complexes through their T cell receptor (TCR). Following TCR-MHC-II-peptide binding, T cells express the surface protein CD40L as well as cytokines such as IL-4 and IL-21. CD40L serves as a necessary co-stimulatory factor for B cell activation by binding the B cell surface receptor CD40, which promotes B cell proliferation, immunoglobulin class switching, and somatic hypermutation as well as sustains T cell growth and differentiation. T cell-derived cytokines bound by B cell cytokine receptors also promote B cell proliferation, immunoglobulin class switching, and somatic hypermutation as well as guide differentiation. After B cells receive these signals, they are considered activated.

Now activated, B cells participate in a two-step differentiation process that yields both short-lived plasmablasts for immediate protection and long-lived plasma cells and memory B cells for persistent protection. The first step, known as the extrafollicular response, occurs outside lymphoid follicles but still in the SLO. During this step activated B cells proliferate, may undergo immunoglobulin class switching, and differentiate into plasmablasts that produce early,

weak antibodies mostly of class IgM. The second step consists of activated B cells entering a lymphoid follicle and forming a germinal center (GC), which is a specialized microenvironment where B cells undergo extensive proliferation, immunoglobulin class switching, and affinity maturation directed by somatic hypermutation. These processes are facilitated by TFH cells within the GC and generate both high-affinity memory B cells and long-lived plasma cells. Resultant plasma cells secrete large amounts of antibody and either stay within the SLO or, more preferentially, migrate to bone marrow.

T cell-independent activation

Antigens that activate B cells without T cell help are known as T cell-independent (TI) antigens and include foreign polysaccharides and unmethylated CpG DNA. They are named as such because they are able to induce a humoral response in organisms that lack T cells. B cell response to these antigens is rapid, though antibodies generated tend to have lower affinity and are less functionally versatile than those generated from T cell-dependent activation.

As with TD antigens, B cells activated by TI antigens need additional signals to complete activation, but instead of receiving them from T cells, they are provided either by recognition and binding of a common microbial constituent to toll-like receptors (TLRs) or by extensive crosslinking of BCRs to repeated epitopes on a bacterial cell. B cells activated by TI antigens go on to proliferate outside lymphoid follicles but still in SLOs (GCs do not form), possibly undergo immunoglobulin class switching, and differentiate into short-lived plasmablasts that produce early, weak antibodies mostly of class IgM, but also some populations of long-lived plasma cells.

Memory B cell activation

Memory B cell activation begins with the detection and binding of their target antigen, which is shared by their parent B cell. Some memory B cells can be activated without T cell help, such as certain virus-specific memory B cells, but others need T cell help. Upon antigen binding, the memory B cell takes up the antigen through receptor-mediated endocytosis, degrades it, and presents it to T cells as peptide pieces in complex with MHC-II molecules on the cell membrane. Memory T helper (TH) cells, typically memory follicular T helper (TFH) cells, that were derived from T cells activated with the same antigen recognize and bind these MHC-II-peptide complexes through their TCR. Following TCR-MHC-II-peptide binding and the relay of other signals from the memory TFH cell, the memory B cell is activated and differentiates either into plasmablasts and plasma cells via an extrafollicular response or enter a germinal center reaction where they generate plasma cells and more memory B cells. It is unclear whether the memory B cells undergo further affinity maturation within these secondary GCs.

Genetic and Molecular Epidemiology Principles

Hardy Weinberg Equilibrium

Hardy Weinberg Equilibrium: In order to believe the result of a linkage analysis, the gene in question needs to be in Hardy Weinberg equilibrium, so in other words, the genotype frequency should only rely on the allele recombination frequency alone (independent probabilities). Otherwise, if the gene was under some selective or external pressures to select one allele over another, that would throw off our expectations about the linkage analysis. You can test that by comparing your observed frequency of haplotype to the expected values using a variation of a chi square test.

The allele frequencies at each generation are obtained by pooling together the alleles from each genotype of the same generation according to the expected contribution from the homozygote and heterozygote genotypes, which are 1 and 1/2, respectively:

$$f_t(\mathrm{A}) = f_t(\mathrm{AA}) + \tfrac{1}{2} f_t(\mathrm{Aa})$$

$$f_t(\mathrm{a}) = f_t(\mathrm{aa}) + \tfrac{1}{2} f_t(\mathrm{Aa})$$

Allele frequency refers to how frequently a particular allele appears in a population. For instance, if all the alleles in a population of pea plants were purple alleles, *W*, the allele frequency of *W* would be 100%, or 1.0. However, if half the alleles were *W* and half were *w*, each allele would have an allele frequency of 50%, or 0.5.

In general, we can define allele frequency as

$$\text{Frequency of allele } A = \frac{\text{Number of copies of allele } A \text{ in population}}{\text{Total number of } A/a \text{ gene copies in population}}$$

Sometimes there are more than two alleles in a population. In that case, you would want to add up *all* of the different alleles to get your denominator.

It's also possible to calculate **genotype frequencies**—the fraction of individuals with a given genotype—and **phenotype frequencies**—the fraction of individuals with a given phenotype. Keep in mind, though, that these are different concepts from allele frequency.

Calculating Allele Frequencies

Genotype	AA	Aa	aa
Frequency	75	20	5

If you have 100 genes, that means that you would have 200 alleles, when looking at allele frequency. In order to calculate the frequency of allele A, the homozygous AA would contribute 75 x2 A alleles, whereas the Heterozygous Aa would contribute 20 x1 A alleles. The A allele frequency would therefore be (75 + 75 + 20)/200 = 0.85

The different ways to form genotypes for the next generation can be shown in a Punnett square, where the proportion of each genotype is equal to the product of the row and column allele frequencies from the current generation.

Table 1: Punnett square for Hardy–Weinberg

		Females	
		A (*p*)	**a (*q*)**
Males	**A (*p*)**	AA (p^2)	Aa (pq)
	a (*q*)	Aa (qp)	aa (q^2)

The sum of the entries is $p^2 + 2pq + q^2 = 1$, as the genotype frequencies must sum to one.

Note again that as $p + q = 1$, the binomial expansion of $(p + q)^2 = p^2 + 2pq + q^2 = 1$ gives the same relationships.

Summing the elements of the Punnett square or the binomial expansion, we obtain the expected genotype proportions among the offspring after a single generation:

$$f_1(\mathrm{AA}) = p^2 = f_0(\mathrm{A})^2$$
$$f_1(\mathrm{Aa}) = pq + qp = 2pq = 2f_0(\mathrm{A})f_0(\mathrm{a})$$
$$f_1(\mathrm{aa}) = q^2 = f_0(\mathrm{a})^2$$

These frequencies define the Hardy–Weinberg equilibrium. It should be mentioned that the genotype frequencies after the first generation need not equal the genotype frequencies from the initial generation, e.g. $f_1(\mathrm{AA}) \neq f_0(\mathrm{AA})$. However, the genotype frequencies for all *future* times will equal the Hardy–Weinberg frequencies, e.g. $f_t(\mathrm{AA}) = f_1(\mathrm{AA})$ for $t > 1$. This follows since the genotype frequencies of the next generation depend only on the allele frequencies of the current generation which, as calculated by equations (**1**) and (**2**), are preserved from the initial generation:

$$f_1(\mathrm{A}) = f_1(\mathrm{AA}) + \tfrac{1}{2}f_1(\mathrm{Aa}) = p^2 + pq = p(p+q) = p = f_0(\mathrm{A})$$
$$f_1(\mathrm{a}) = f_1(\mathrm{aa}) + \tfrac{1}{2}f_1(\mathrm{Aa}) = q^2 + pq = q(p+q) = q = f_0(\mathrm{a})$$

To test whether your allele frequencies are in HW, you can compare your punnet square based on your data (observed), to the HW punnet square (expected), and perform a chi-square test:

Remember the basic formulas:

$p^2 + 2pq + q^2 = 1$ *and* $p + q = 1$

p = frequency of the dominant allele in the population
q = frequency of the recessive allele in the population
p^2 = percentage of homozygous dominant individuals
q^2 = percentage of homozygous recessive individuals
2pq = percentage of heterozygous individuals

Pearson's Chi-square:

X^2 = ∑[observed – expected]^2/expected

X^2 = [(Observed p – p^2)^2/p^2] + [(Observed pq – 2pq)^2/2pq] + [(Observed q – q^2)^2/q^2]

Get the p-value based on the X^2 (3.84 is the critical X^2 value for a p<0.05

Genetic Epidemiologic Studies

SNPs

Single nucleotide polymorphisms, frequently called SNPs (pronounced “snips”), are the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called a nucleotide. For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA.

SNPs occur normally throughout a person’s DNA. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome. Most commonly, these variations are found in the DNA between genes. They can act as biological markers, helping scientists locate genes that are associated with disease. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene’s function.

Most SNPs have no effect on health or development. Some of these genetic differences, however, have proven to be very important in the study of human health. Researchers have found SNPs that may help predict an individual’s response to certain drugs, susceptibility to environmental factors such as toxins, and risk of developing particular diseases. SNPs can also be used to track the inheritance of disease genes within families. Future studies will work to identify SNPs associated with complex diseases such as heart disease, diabetes, and cancer.

Candidate Gene Approach

The candidate gene approach to conducting genetic association studies focuses on associations between genetic variation within pre-specified genes of interest and phenotypes or disease states. This contrasts with genome-wide association studies (GWAS), which scan the entire genome for common genetic variation. Candidate genes are most often selected for study based on a priori knowledge of the gene's biological functional impact on the trait or disease in question. The rationale behind focusing on allelic variation in specific, biologically relevant regions of the genome is that certain mutations will directly impact the function of the gene in question, and lead to the phenotype or disease state being investigated. This approach usually uses the case-control study design to try to answer the question, "Is one allele of a candidate gene more frequently seen in subjects with the disease than in subjects without the disease?"

Genome-Wide Association Study

A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

With the completion of the Human Genome Project in 2003 and the International HapMap Project in 2005, researchers now have a set of research tools that make it possible to find the genetic contributions to common diseases. The tools include computerized databases that contain the reference human genome sequence, a map of human genetic variation and a set of new

technologies that can quickly and accurately analyze whole-genome samples for genetic variations that contribute to the onset of a disease.

To carry out a genome-wide association study, researchers use two groups of participants: people with the disease being studied and similar people without the disease. Researchers obtain DNA from each participant, usually by drawing a blood sample or by rubbing a cotton swab along the inside of the mouth to harvest cells. Each person's complete set of DNA, or genome, is then purified from the blood or cells, placed on tiny chips and scanned on automated laboratory machines. The machines quickly survey each participant's genome for strategically selected markers of genetic variation, which are called single nucleotide polymorphisms, or SNPs.

If certain genetic variations are found to be significantly more frequent in people with the disease compared to people without disease, the variations are said to be "associated" with the disease. The associated genetic variations can serve as powerful pointers to the region of the human genome where the disease-causing problem resides.

However, the associated variants themselves may not directly cause the disease. They may just be "tagging along" with the actual causal variants. For this reason, researchers often need to take additional steps, such as sequencing DNA base pairs in that particular region of the genome, to identify the exact genetic change involved in the disease. The National Center for Biotechnology Information (NCBI), a part of NIH's National Library of Medicine, is developing databases for use by the research community. An archive of data from genome-wide association studies on a variety of diseases and conditions already can be accessed through an NCBI Web site, called the Database of Genotype and Phenotype (dbGaP)

Whole Exome Sequencing

Determining the order of DNA building blocks (nucleotides) in an individual's genetic code, called DNA sequencing, has advanced the study of genetics and is one technique used to test for genetic disorders. Two methods, whole exome sequencing and whole genome sequencing, are increasingly used in healthcare and research to identify genetic variations; both methods rely on new technologies that allow rapid sequencing of large amounts of DNA. These approaches are known as next-generation sequencing (or next-gen sequencing).

With next-generation sequencing, it is now feasible to sequence large amounts of DNA, for instance all the pieces of an individual's DNA that provide instructions for making proteins. These pieces, called exons, are thought to make up 1 percent of a person's genome. Together, all the exons in a genome are known as the exome, and the method of sequencing them is known as whole exome sequencing. This method allows variations in the protein-coding region of any gene to be identified, rather than in only a select few genes. Because most known mutations that cause disease occur in exons, whole exome sequencing is thought to be an efficient method to identify possible disease-causing mutations.

The exome is the part of the genome formed by exons, the sequences which when transcribed remain within the mature RNA after introns are removed by RNA splicing. It consists of all DNA that is transcribed into mature RNA in cells of any type as distinct from the transcriptome, which is the RNA that has been transcribed only in a specific cell population. The exome of the human genome consists of roughly 180,000 exons constituting about 1% of the total genome, or about 30 megabases of DNA. Though comprising a very small fraction of the genome, mutations

in the exome are thought to harbor 85% of mutations that have a large effect on disease. Exome sequencing has proved to be an efficient strategy to determine the genetic basis of more than two dozen Mendelian or single gene disorders.

However, researchers have found that DNA variations outside the exons can affect gene activity and protein production and lead to genetic disorders--variations that whole exome sequencing would miss. Another method, called whole genome sequencing, determines the order of all the nucleotides in an individual's DNA and can determine variations in any part of the genome.

While many more genetic changes can be identified with whole exome and whole genome sequencing than with select gene sequencing, the significance of much of this information is unknown. Because not all genetic changes affect health, it is difficult to know whether identified variants are involved in the condition of interest. Sometimes, an identified variant is associated with a different genetic disorder that has not yet been diagnosed (these are called incidental or secondary findings).

Epigenetics Quick Review

The primary interest of some genetic studies is to identify the primordial factor on which we can intervene to reduce risk. A lot of the factors we are interested in, such as transgenerational obesity can be due to either a gene x environment interaction (compounding the feeding and physical activity tendencies with genetic predisposition), but these can also be affected by the epigenome which can alter gene expression as early as in utero, to the microbiome and its relationship with our immune system, given that biomarkers such as CRP have both inflammatory and antimicrobial properties. In recent years, we have observed that the decline in infection disease as been matched with a steady increase in chronic disease. This pattern is known as the hygiene hypothesis, and illustrates how changes in our environment have led to changes in our health. Our microbiome and epigenome are the key mechanisms through which these changes have occurred.

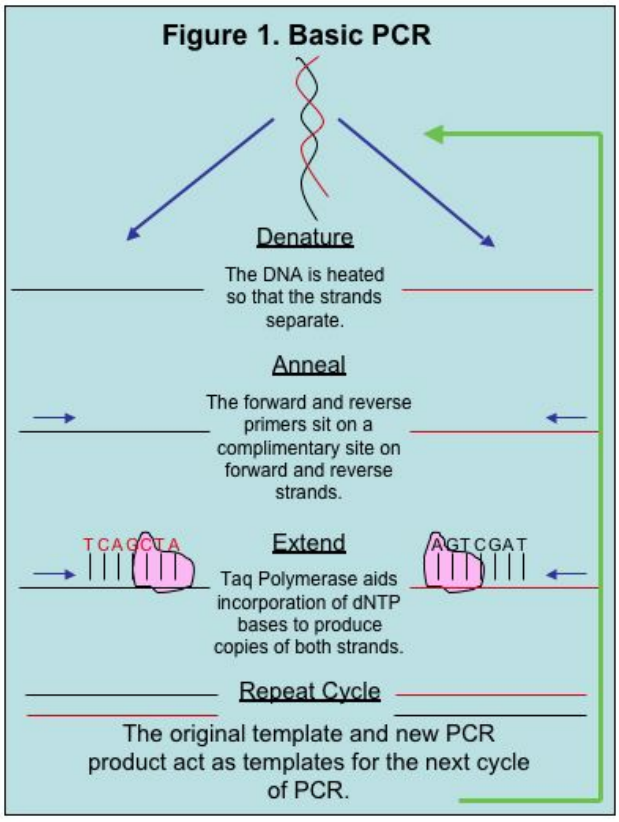
Multiplex PCR

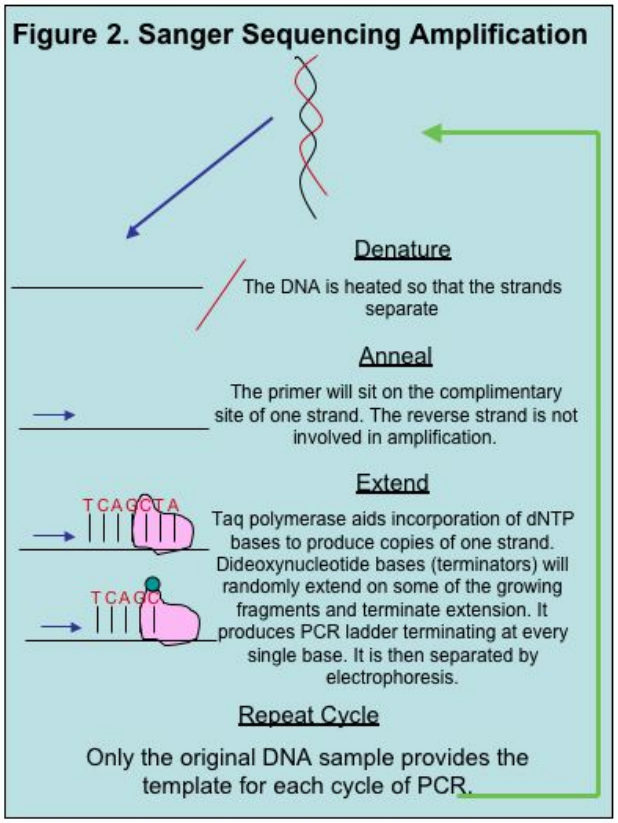
Introduction of Multiplex PCR

Multiplex PCR is a widespread molecular biology technique for amplification of multiple targets in a single PCR experiment. In a multiplexing assay, more than one target sequence can be amplified by using multiple primer pairs in a reaction mixture. As an extension to the practical use of PCR, this technique has the potential to produce considerable savings in time and effort within the laboratory without compromising on the utility of the experiment.

Multiplexing reactions can be broadly divided in 1) Single Template PCR Reaction (This technique uses a single template which can be a genomic DNA along with several pairs of forward and reverse primers to amplify specific regions within a template.) and 2) Multiple Template PCR Reaction (It uses multiple templates and several primer sets in the same reaction tube. Presence of multiple primers may lead to cross hybridization with each other and the possibility of mis-priming with other templates.)

Sanger sequencing, the process used for automated sequencing, requires a DNA template to be amplified by the Polymerase Chain Reaction (PCR). Despite similarities between the processes, a sequencing amplification is different than basic PCR.





Sanger sequencing utilizes linear amplification

PCR produces millions of copies of a DNA region from a single copy of template DNA. Each copy produced during PCR in one cycle becomes a new template for the next cycle. PCR uses forward and reverse primers. The forward primer anneals to a complimentary site on one strand of DNA and extends toward the reverse primer. In turn, the reverse primer similarly extends towards the forward primer. What results is a copy of the desired region of DNA to be amplified. The new copy contains priming sites so it can be used as a template for future amplifications (figure 1). One copy of the original template produces two copies; two copies produce four in the next cycle; and so on. A twenty-five cycle PCR will produce 2E24 copies from a single template.

Sanger sequencing uses one primer instead of two. The amplification process copies one strand but not the reverse strand. The copy is the same direction as the primer and cannot be used as a template for later cycles. All amplification is directly from the original template DNA in the reaction. Therefore, amplification is linear, not exponential. It is the reason that Sanger sequencing amplification must include sufficient copies of the original template DNA to be visualized by automated sequencing equipment (figure 2).

Dideoxynucleotide bases are included in Sanger sequencing

The components of basic PCR include buffer, the enzyme Taq polymerase, deoxynucleotides (dNTPs), template, and forward and reverse primers. Sanger sequencing includes an additional component called dideoxynulceotides (ddNTPs). The ddNTPs are terminating bases that include a fluorescent tag for automated sequencing equipment. For this reason, Sanger sequencing is also called dye-terminator sequencing. During amplification, the ddNTPs will randomly sit on the DNA template and terminate the extension. The dNTPs sit on the remaining templates and continue extending. The end product is a size ladder of PCR products that increase by a single base (figure 2). Each terminating base is tagged with fluorescent dye. This dye provides a unique color representing the A, G, C, and T bases in DNA.

The DNA ladder is separated by electrophoresis

Once PCR is complete, the products produced in the Sanger reaction are loaded on an automated slab gel or capillary analyzer. Products will separate by size with smaller products moving faster through the medium. As the products near the end of the medium, the fluorescent tags are excited by light and recorded to a computer with a digital camera (figure 3). The computer records the color for each band and assigns the correct base to complete dye-terminator sequencing.

Capillary Electrophoresis Sequencing

In capillary sequencing machines, DNA fragments are separated by size through a long, thin, acrylic-fibre capillary (instead of an electrophoresis gel, as with the Sanger method). A sample containing fragments of DNA is injected into the capillary. This is done by dipping the capillary and an electrode into a solution of the sample, and briefly applying an electric current. This causes the DNA fragments to migrate on to the end of the capillary.

Once the sample has been injected, the electric field is reapplied, to drive the DNA fragments through the capillary. A fluorescence-detecting laser, built into the machine, then shoots through the capillary fibre, causing the coloured tags on the DNA fragments, to fluoresce. Each base terminator is labelled with a different colour: A = Green, C = Blue, G = Yellow and T = Red.

The colour of the fluorescent bases is detected by a camera, and recorded by the sequencing machine.The colours of the bases are then displayed on a computer as a graph of different coloured peaks. The small diameter of the capillary allows for the use of extremely high electric fields, and consequently, very rapid separation of DNA sequencing fragments.

Basics of NGS Chemistry

In principle, the concept behind NGS technology is similar to CE sequencing. DNA polymerase catalyzes the incorporation of fluorescently labeled deoxyribonucleotidetriphosphates (dNTPs) into a DNA template strand during sequential cycles of DNA synthesis. During each cycle, at the point of incorporation, the nucleotides are identified by fluorophore excitation. The critical difference is that, instead of sequencing a single DNA fragment, NGS extends this process across millions of fragments in a massively parallel fashion. More than 90% of the world's sequencing data are generated by Illumina sequencing by synthesis (SBS) chemistry. It delivers high accuracy, a high yield of error-free reads, and a high percentage of base calls aboveQ30.

Illumina NGS workflows include four basic steps:

1. LibraryPreparation: The sequencing library is prepared by random fragmentation of the DNA or cDNA sample, followed by 5′ and 3′ adapter ligation(Figure3A). Alternatively, “tagmentation” combines the fragmentation and ligation reactions into a single step that greatly increases the efficiency of the library preparation process. Adapter-ligated fragments are then PCR amplified and gel purified.

2. Cluster Generation: For cluster generation, the library is loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters. Each fragment is then amplified into distinct, clonal clusters through bridge amplification(Figure3B). When cluster generation is complete, the templates are ready for sequencing.

3. Sequencing: Illumina SBS technology uses a proprietary reversible terminator–based method that detects single bases as they are incorporated into DNA template strands(Figure3C). As all four reversible terminator–bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias and greatly reduces raw error rates compared to other technologies. The result is highly accurate base-by-base sequencing that virtually eliminates sequence context–specific errors, even within repetitive sequence regions and homopolymers.

4. Data Analysis: During data analysis and alignment, the newly identified sequence reads are aligned to a reference genome (Figure3D). Following alignment, many variations of analysis are possible, such as single nucleotide polymorphism(SNP) or insertion-deletion(indel) identification, read counting for RNA methods, phylogenetic or metagenomic analysis, and more.

Microbiomics Quick Review

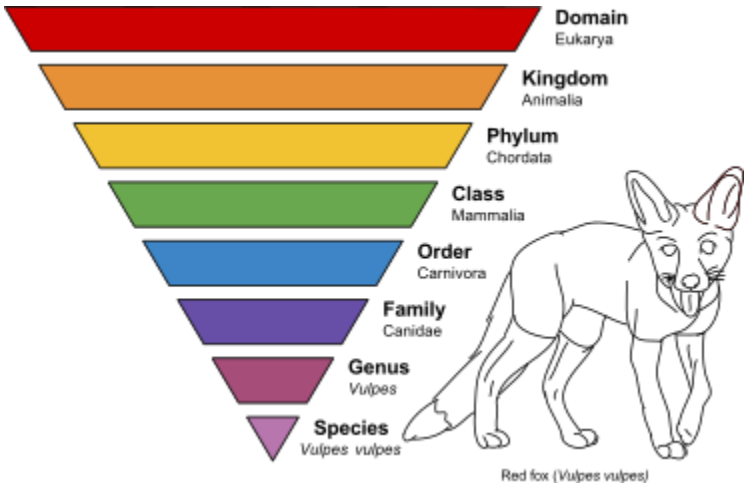
The initial idea behind microbiomics is to sequence the hypervariable sequence of the 16s gene (common to most microbiome species, as it is part of the gene coding for the bacterial mitochondria) and use these reads to get relative quantities of bacteria, and associate them to a specific genus or species. However, one problem (specifically using the microarray probe approach, such as the HOMIM) is that too many species and even genuses share a large proportion on their genome, so the different probes will correspond to the same organism (because any given probe will interrogate multiple corresponding species)

The first improvement to this method was made when scientists used a pyrosequencing approach with 454s sequencing, which is nowadays done with the Illumina platform. For each sample you sequence, you will come up with a certain number of counts that identify a sequence to a genus and species (matched within a database, either the RGP (which can have many false positives) or UPARSE (which is more conservative). The count approach can be time consuming, so it is preferable to first cluster similar sequences together (similar to using a k means approach), come up with a concensus sequence (which becomes your OTU) , then map that information to an organism from a database.

Pipelines in R are packages designed to contain all of the functions you would need to run this analysis using next generation sequencing. The QIIME pipeline needs a lot of updates to make sure that all the individual programs are in sync, whereas the MOTHUR pipeline rewrites the packages to make sure that they are in sync. These approaches use the EdgeR or DESeq approach in R to run the analysis, and identify species within the population that are overrepresented in either disease condition.

Assigning Taxonomy

Once you have clustered your sequences, you can use a reference database to assign various taxonomic ranks to the OTU clusters you have created



In QIIME, L2=phylum, L3 = Class, L4 = Order, L5 = Family, L6 = Genus, L7 = Species

Files needed for a microbiome analysis

Raw Sequencing Data (.sff) for the Denoiser

This Denoiser step only applies to 454 Pyrosequencing data. It can be skipped, at your own risk. One problem with 454 pyrosequencing is that sequencing errors can give you effectively more OTUs than really are there. There are a number of strategies for dealing with this problem. The default in QIIME is to use a built-in program called Denoiser, which compares flowgrams (the raw sequencing data) of similar sequences to see if the differences between them may have been due to erroneous base calls. You can see some of this raw sequencing flowgram data by opening the Fasting_Example.sff.txt file in less. This is a text version of the "SFF" raw data format generated by the 454 pyrosequencing platform. It won't be terribly meaningful to you as it is, but it's always nice to have a feel for what all the data files look like.

Sequences (.fna)

This is the 454-machine generated FASTA file. Using the Amplicon processing software on the 454 FLX standard, each region of the PTP plate will yield a fasta file of form 1.TCA.454Reads.fna, where "1" is replaced with the appropriate region number.

The primary file format for storing sequence data supported by QIIME is the FASTA format. The file Fasting_Example.fna is in FASTA format, indicated by the suffix ".fna" which stands for **F**ASTA **n**ucleic **a**cids (as opposed to amino acids which would have a suffix ".faa").

Quality Scores (.qual)

This is the 454-machine generated quality score file, which contains a score for each base in each sequence included in the FASTA file. Like the fasta file mentioned above, the Amplicon processing software will generate one of these files for each region of the PTP plate, named 1.TCA.454Reads.qual, etc. For the purposes of this tutorial, we will use the quality scores file Fasting_Example.qual.

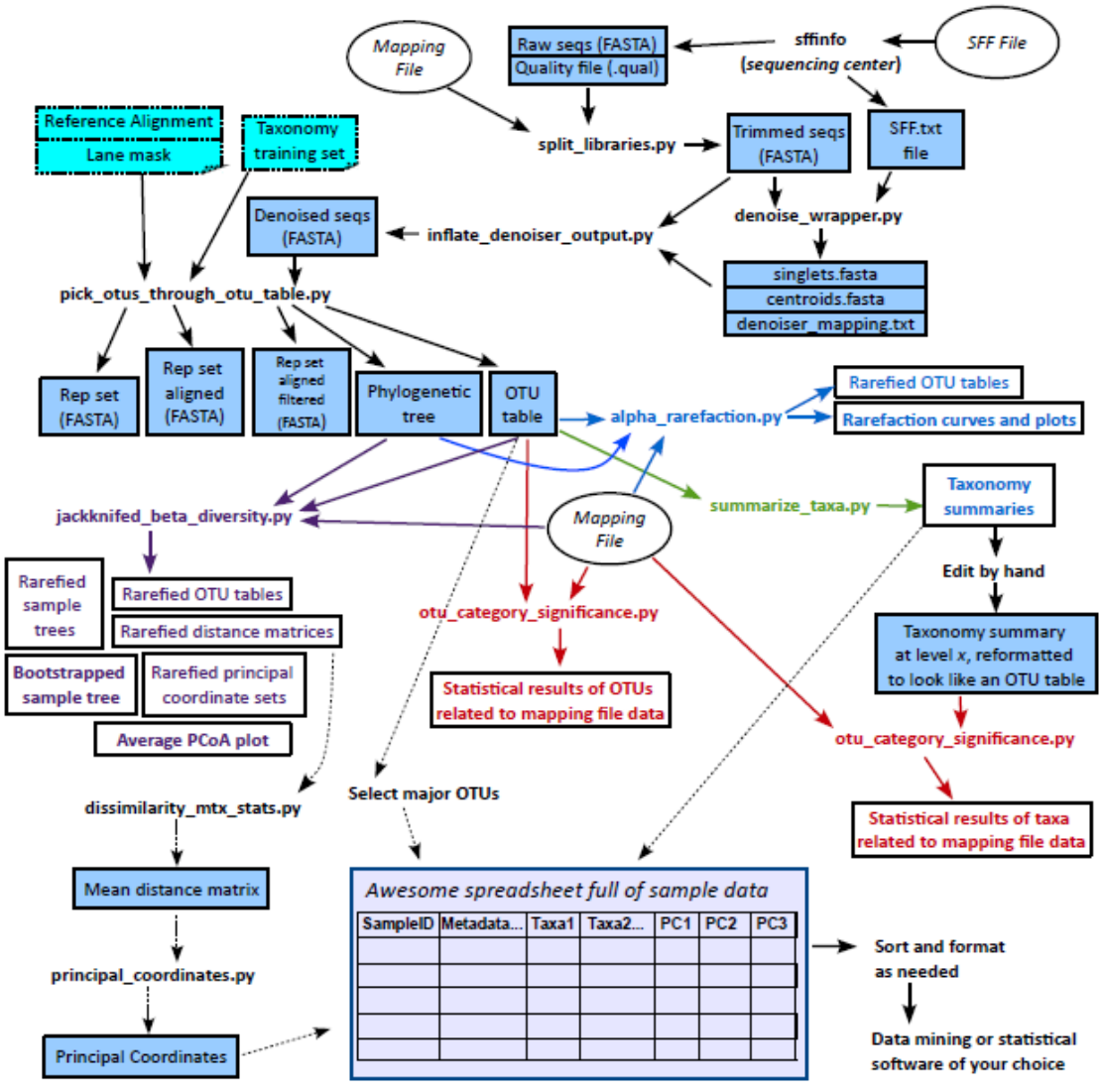
Mapping File (Tab-delimited .txt)

The mapping file is generated by the user. This file contains all of the information about the samples necessary to perform the data analysis.

An example of a sequence analysis pipeline in QIIME

Werner | March 7, 2012

This is just an example of a good way to get started with a 16S rRNA gene pyrosequencing survey. The two things you start out with, after sequencing is completed, are the SFF file (sequencing data) and your mapping file (your experimental data about the samples you submitted for sequencing). Hopefully the sequencing center will also send you the fasta, qual, and sff.txt files. One thing you may want to insert in there would be chimera checking with identify_chimeric_seqs.py - input files would be the "Rep set aligned (FASTA)" file (not the pfiltered one) and the reference alignment. Once you've identified chimeric seqs, all you have to do is delete them from the OTU table.



Bioinformatics applications of metabolic analyses

- Genomics
- Epigenetics
- Transcriptomics
- Proteomics
- Metabolomics

Proteomics and metabolomics can both be assessed using either GS (gas) or LC (liquid) chromatography for basic and translational research. We can use either untargeted (qualitative, where we determine the relative amount of expression) or targeted (quantitative) to assess the metabolome in relation or a phenotype (proteomics) or disease states. We can also assess the concentration or the rate of appearance/disappearance of metabolites (fluxomics), as both affect phenotype or disease states.

Modeling with phenotype or metabolite levels.

We can use targeted or untargeted metabolomics to determine relations of the metabolome to proteomics or gene expression because metabolite levels are driven by enzymatic reactions, and enzymes are made of proteins. Therefore, we can look at:

- Metabolome
- Proteomics + Metabolome
- Gene Expression + Proteomics + Metabolome

LS-MS Metabolomics workflow

Pre-separation of polar (CL8 column) and non-polar (HILIC colum) compounds in chromatography column to allow the compounds to be retained long enough for the mass spectrophotometer to detect them.

Further separation of compound in CL8 positive / CL8 negative / HILIC positive / HILIC negative columns, as different compounds have different retention times based on the charge they carry when they are in the mass spectrophotometer. Always run biological replicates (multiple samples per group) and technical replicates (multiple injections per samples). A rule of thumb is to run 3 technical replicates to minimize the chance for false positive results, as it is usually costlier to try to ID a new compound than it is to run a technical replicate.

Normalize the 4 resulting datasets you obtain from the previous step. You can either combine the 4 modes (CL8 positive / CL8 negative / HILIC positive / HILIC negative) across samples (ie, combine all the CL8 +, CL8 – ect together, so you get 1 reading per mode, with all your samples), then normalize them to a single master dataset, or you can combine all samples across the 4 modes (ie you get 1 reading per patient, with all 4 modes at once). The benefits of the former is that although you may get redundant results, with complementary analyses, it is rare, and further you can address them by using Fisher's exact test to average the values before transforming the data).

Differential analysis: Here you will go through feature detection (reads peaks which define a compound), RT alignment (overlap the peaks from different samples) to account for systematic error (shift in peak location) Intensity transformation (determined by calculating the AUC under each peak to tell us how much of a compound is present), multisampling normalization (log transformation of the data to get an approximately normal data distribution) and metabolite ID assignment (match compound to a database based on the peak intensities: these can be tier 1 IDs based on synthetic compounds & reference standards, Tier 2 IDs based on mass spectra library matching, or tier 3 based on the combination of metabolite mass and isotopic distribution). This step yields a file with compound IDs, normalized & non normalized intensities fold change, P values and corrected p-values. This file can them be used for statistical and pathways analyses.

Pathway analysis: The first approach is to visualize the data using principle component analysis. This method determines the combination of metabolites that best explain the largest amount of variation between the group (i.e. treatment groups or disease states). PCA is useful if your dependent variable is a dicothomous variable (disease state, phenotype, or treatment group), but you will need to use regression if you have a quantitative predictor. The second approach is PLS-DA, which allows us to rank compounds based on their optimal order for differentiating groups (again, groups being a dicothomous variable such as treatment group or disease state). In your analyses, take the metabolites that were consistently identified at a high (3^{rd} or 4^{th} intensity quartile) across your technical replicate).

Impact of increasing the sample size on your p-value: p-value give you an idea of seeing a signal / noise ratio as large or more extreme as the one you observed in your experiment. Signal / noise can be interpreted as fold change intensity / standard deviation of the metabolites, so if you have a small standard deviation (which metabolites usually do) then you will have a large signal to noise ratio, and thus a small p-value. Given that standard deviation = $\sqrt{\sigma}/n$, when we increase N, we get a smaller standard deviation, thus improving our p-value.

Differences between NMR and MS metabolomics: both approaches are used to relate analytes to phenotypes or disease, but with NMR, we have less variability due to systematic error, and the sample are easier to prep, and we do not have to worry about batch effects.

NMR analysis: the solvent choice has a big impact on the proton nmr, and is also used to create a single reference point peak (you can also use the commercial ERETIC digital reference signal). You will need a S/N ratio of at least 10 to reliably quantify a compound. The signal to noise ration improves at a 2^n rate, thus to get a 2x improvement on a spectrum, we need to scan it 2^2 = 4 times. Biological compounds typically require 64/128 scans for good accuracy. The quantification threshold is at the micromolar (10^-6) range. NMR is better suited for targeted analysis, whereas GS can be used for untargeted metabolomics. When you have multiple compounds, you can use NMR bins to reduce the analysis to a peak position + intensity, which allows you to minimize issues with sample dilution (which causes drifts in the peak location). A wide bin will yield less info, and a slim bin will yield more information on a compound.

Software: We can use the XCMS free software to load our MS data (compound ID, FC, p-value) and conduct integrated pathways analyses. We can also use the IPMala software to combine

differentially expressed gene expression data (proteomics, genomics) with metabolomics data. By using both the differentially expressed gene and differentially expressed pathways, we get a better picture of the specific pathways which are targeted by a treatment or disease state, or phenotype. The Chenomix NMR software offer similar capacities as the XCMS MS software. With the Chenomix software, we can either upload a spectrum to find the matching metabolite, or we can pick metabolites a priori and scan the spectra from our sample to determine their relative quantities.

Principal Component vs Principal Coordinates Analysis

Principal Component Analysis and Principal Coordinates Analysis cover two of the main mathematical approaches to multivariate analyses. I think the main use of these methods is to visualize the data, but more complicated analyses can be done using the same ideas (e.g. MANOVA and factor analysis are based on the PCA approach).

Principal Component Analysis

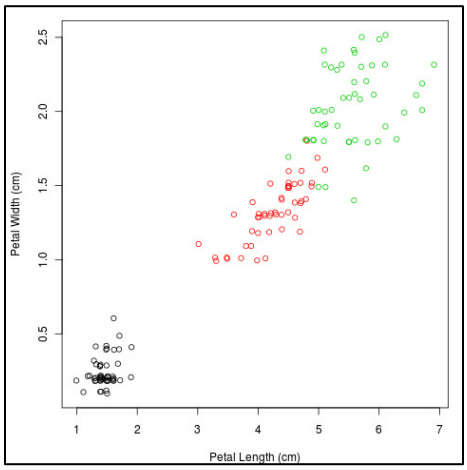
- Sometimes, we have variables in our dataset that are highly related to one another.
- This redundancy leads to correlation which creates unneeded noise in our analysis
- With PCA, we take all of these redundant variables, and make new index variables that get rid of the redundancy by grouping the more similar variables together
- Therefore, PCA is an analytical tool to find patterns in our data using their variance/covariance. It compresses the larger dimensions into similar groups without loosing information by using linear combinations: each variable in the combination gets its own weight depending on how much it contributes to the groupiness of the new index variable.

Criteria to determine the number of components vary: It can either be based on the eigenvalue (greater than 1), the scree test (visual version of the eigen value)

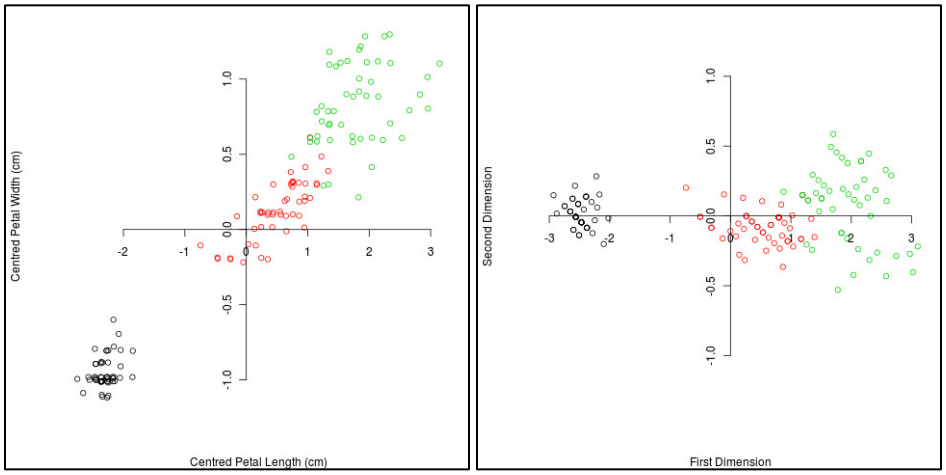
- Step 1: calculate the correlation between all the variables you want to use in the component analysis (to get a sense of the grouping)
- Step 2: run the PCA extraction in the variables
- Step3: look at the eigen values, or multiple class modeling based on AIC/BIC to determine the ideal number of classes
- Step 4: create the component scores, and check their correlation: there should be none since the redundancy was removed.
- The gamma parameter in PCA (LCA) gives you the proportion of the sample in each class, and the rho parameter gives you the proportion of each variable in your dataset that endorses the class you're looking at (variables that score more than 75% in one class should be the ones grouped together)

Principal Component Analysis: PCA

PCA is a statistical yoga warm-up: it's all about stretching and rotating the data. I'll illustrate it with part of a famous data set, of the size and shape of iris flowers. The original data has 4 dimensions: sepal and petal length and width. Here are the petal measurements (the different colours are different species):



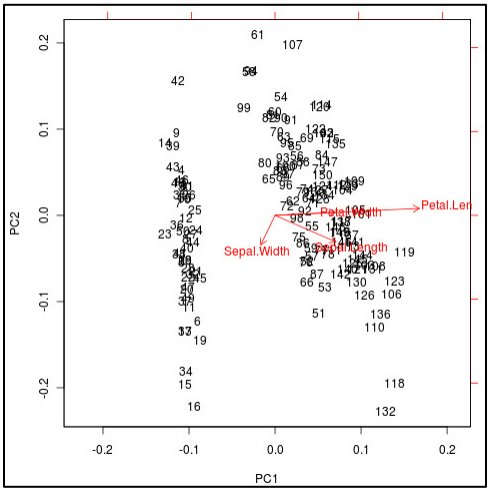
The purpose of PCA is to represent as much of the variation as possible in the first few axes. To do this we first center the variables to have a mean of zero, and then rotate the data (or rotate the axes, but I can't work out how to do that in R):



The rotation is done so that the first axis contains as much variation as possible, the second axis contains as much of the remaining variation etc. Thus if we plot the first two axes, we know that these contain as much of the variation as possible in 2 dimensions. As well as rotating the axes, PCA also re-scales them: the amount of re-scaling depends on the variation along the axis. This can be measured by the Eigenvalue, and it's common to present the proportion of total variation

as the Eigenvalue divided by the sum of the Eigenvalues, e.g. for the data above the first dimension contains 99% of the total variation.

We can plot these on a graph: here we have it for the full data, with both sepals and petals:

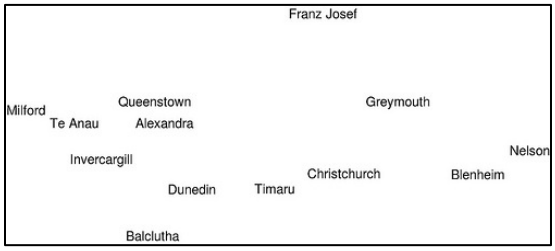


The arrows show the direction the variables point, so we can see that the petal variables are pretty much the same (i.e. there is a petal size which affects both equally). Sepal width is the only one that's different, mainly affecting the second PC. It's almost at 90° from the sepal length, suggesting they have independent effects.

It's obvious, looking at the data, that one species (I. setosa) is very different, with smaller petals, and differently shaped sepals. Mathematically, PCA is just an eigen analysis: the covariance (or correlation) matrix is decomposed into its Eigenvectors and Eigenvalues. The Eigenvectors are the rotations to the new axes, and the Eigenvalues are the amount of stretching that needs to be done. But you didn't want to know that, did you?

Principal Coordinate Analysis: PCoA

The second method takes a different approach, one based on distance. For example, we can take a map of towns in New Zealand:



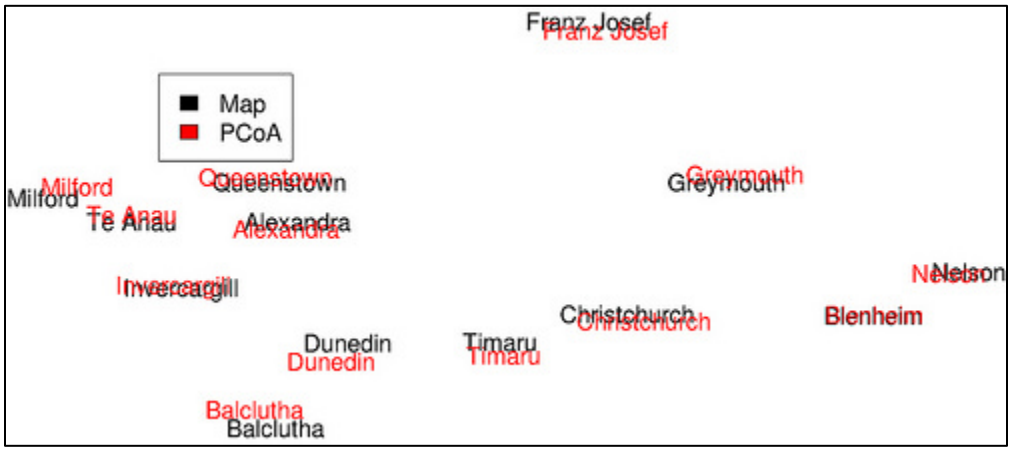
(I've rotated the map, to make it easier later)

The geographical distance between the town is Euclidean (well, almost. And near enough that you won't notice the difference). But we can also measure the distance one would travel by road between the towns. This would give us another measure of distance. Of course this may not be Euclidean in two dimensions, so we can't simply plot it onto a piece of paper. Distance-based

methods are essentially about finding a “good” set of Euclidean distances from distances that are not a priori Euclidean in those dimensions.

The way principal coordinate analysis does this is to start off by projecting the distances into Euclidean space in a larger number of dimensions. This is not difficult; as long as the distances are fairly well behaved then we only need n-1 dimensions for with n data point. PCoA starts by putting the first point at the origin, and the second along the first axis the correct distance from the first point, then adds the third so that the distance to the first 2 is correct: this usually means adding a second axis. This continues until all of the points are added.

But how do we get back down to 2 dimensions? Well, simply do a PCA on these constructed points. This obviously captures the largest amount of variation from the n-1 dimensional space. So, for the New Zealand data if we do PCoA on the road distances, we get this:



And, not surprisingly, the maps are fairly close to each other, but not exact.One wrinkle for the sorts of applications we were discussing for bioinformatics (and which is also important in ecology) is the notion of a distance between two data points. In ecology we work with abundances of species, and the distances between the points need to somehow scale the abundances, so that rarer and more common species. So a plethora of diversity indices have been devised, and it’s not clear which one should be used (unless one is in Canberra or New York, I guess).

For me, these methods are mainly useful for visualizing the data, so we can actually see how it’s behaving, and they’re not really very good for making formal inferences. But a lot of the multivariate methods for inference have the same ideas at their core, so understanding these is a good starting point. But it’s still all about avoiding headaches by not having to think in 17 dimensions

Biostatistics Review

Natural Logarithms

Natural Logarithm Rules & Properties

Rule name	Rule	Example
Product rule	$\ln(x \cdot y) = \ln(x) + \ln(y)$	$\ln(3 \cdot 7) = \ln(3) + \ln(7)$
Quotient rule	$\ln(x / y) = \ln(x) - \ln(y)$	$\ln(3 / 7) = \ln(3) - \ln(7)$
Power rule	$\ln(x^y) = y \cdot \ln(x)$	$\ln(2^8) = 8 \cdot \ln(2)$
Ln derivative	$f(x) = \ln(x) \Rightarrow f'(x) = 1 / x$	
Ln integral	$\int \ln(x)dx = x \cdot (\ln(x) - 1) + C$	
Ln of negative number	$\ln(x)$ is undefined when $x \leq 0$	
Ln of zero	$\ln(0)$ is undefined	
	$\lim_{x \to 0^+} \ln(x) = -\infty$	
Ln of one	$\ln(1) = 0$	
Ln of infinity	$\lim \ln(x) = \infty$, when $x \to \infty$	

Derivative of natural logarithm (ln) function

The derivative of the natural logarithm function is the reciprocal function.

When

$$f(x) = \ln(x)$$

The derivative of f(x) is:

$$f'(x) = 1 / x$$

What does ln(0) = ?

What is the natural logarithm of zero?

$$\ln(0) = ?$$

The real natural logarithm function ln(x) is defined only for x>0.

So the natural logarithm of zero is undefined.

$$\ln(0) \text{ is undefined}$$

Why the natural logarithm of zero is undefined?

Since ln(0) is the number we should raise e to get 0:

$$e^x = 0$$

There is no number x to satisfy this equation.

Limit of the natural logarithm of zero

The limit of the natural logarithm of x when x approaches zero from the positive side (0+) is minus infinity:

$$\lim_{x \to 0^+} \ln(x) = -\infty$$

What is the natural logarithm of 1?

What is the natural logarithm of one.

$$\ln(1) = ?$$

The natural logarithm of a number x is defined as the base e logarithm of x:

$$\ln(x) = \log_e(x)$$

So

$$\ln(1) = \log_e(1)$$

Which is the number we should raise e to get 1.

$$e^0 = 1$$

So the natural logarithm of one is zero:

$$\ln(1) = \log_e(1) = 0$$

What is the natural logarithm of e?

What the natural logarithm of the e constant (Euler's constant)?

$$\ln(e) = ?$$

The natural logarithm of a number x is defined as the base e logarithm of x:

$$\ln(x) = \log_e(x)$$

So the natural logarithm of e is the base e logarithm of e:

$$\ln(e) = \log_e(e)$$

ln(e) is the number we should raise e to get e.

$$e^1 = e$$

So the natural logarithm of e is equal to one.

$$\ln(e) = \log_e(e) = 1$$

Natural Logarithm of Negative Number

What is the natural logarithm of a negative number?

The natural logarithm function ln(x) is defined only for x>0.

So the natural logarithm of a negative number is undefined.

$$\ln(x) \text{ is undefined for } x \leq 0$$

The complex logarithmic function Log(z) is defined for negative numbers too.

For $z=r \cdot e^{i\theta}$, the complex logarithmic function:

$$\text{Log}(z) = \ln(r) + i\theta, \quad r > 0$$

So for real negative number $\theta = -\pi$:

$$\text{Log}(z) = \ln(r) - i\pi, \quad r > 0$$

Definitions

Homoscedasticity: The assumption of homoscedasticity (meaning "same variance") is central to linear regression models. Heteroscedasticity (the violation of homoscedasticity) is present when the size of the error term differs across values of an independent variable.

Ecological fallacy: you cannot use a correlation or association observed at the group level to make an inference at the individual level or vice versa. Even though this may be mathematically correct (the average is a good estimate of individual values, it is not practical for making inferences about the individuals, because you are only technically right 1% of the time (when the true value is the average) but wrong 99% of the time (the rest of the distribution)

68-95-99.7 Rule for Normal Distributions

• 68% of the AUC falls within $\pm 1\sigma$ of μ

• 95% of the AUC falls within $\pm 2\sigma$ of μ

• 99.7% of the AUC falls within $\pm 3\sigma$ of μ

Re-expression of Non-Normal Variables

• Many biostatistical variables are not normal, but we can re-express non-Normal variables with a mathematical transformation to make them normal

• Example of mathematical transforms include logarithms, exponents, square roots.

Logarithms are exponents of their base, and there are two main logarithmic bases

– common log10 (base 10)

– natural ln (base e)

Determining Normal Probabilities

To determine a Normal probability when the value does not fall directly on a $\pm 1\sigma$, $\pm 2\sigma$, or $\pm 3\sigma$ landmark, follow this procedure:

1. State the problem

2. Standardize the value (z score)

3. Sketch and shade the curve

4. Use Table B to determine the probability

Standard Normal variable ≡ a Normal random variable with μ = 0 and σ = 0

• Called "Z variables"

• Notation: Z ~ N(0,1)

• To standardize, subtract μ and divide by σ.

• The z-score tells you how the number of σ- units the value falls above or below μ

• For example, the value 40 from X ~ N(39,2) has Z = (40-39)/2 = 0.5

Calculating Probabilities on a curve

Let a represent the lower boundary and b represent the upper boundary of a range:

$Pr(a \leq Z \leq b) = Pr(Z \leq b) - Pr(Z \leq a)$

Review: Basics of Inference

• Population ≡ all possible values

• Sample ≡ a portion of the population

• Statistical inference ≡ generalizing from a sample to a population with calculated degree of certainty

• Two forms of statistical inference – Hypothesis testing – Estimation

Parameter ≡ a numerical characteristic of a population, e.g., population mean μ, population proportion p

Statistic ≡ a calculated value from data in the sample, e.g., sample mean ( ), sample proportion ( )

Hypothesis Testing Steps

1. Null and alternative hypotheses: Convert the research question to null and alternative hypotheses • The null hypothesis (H0) is a claim of "no difference in the population" • The alternative hypothesis (Ha) is a claim of "difference" • Seek evidence against H0 as a way of bolstering Ha

2. Test statistic: Z = xbar – u / (SExbar)

3. P-value and interpretation: The P-value answer the question: What is the probability of the observed a test statistic equal to or one more extreme than the current statistic assuming H0 is true? • This corresponds to the AUC in the tail of the Z sampling distribution beyond the z-stat. Use Table B or a software utility to find this AUC. Smaller and smaller P-values provider stronger and stronger evidence against H0

4. Significance level (optional): Smaller and smaller P-values provide stronger and stronger evidence against H0. Although it unwise to draw firm cutoffs, here are conventions that used as a starting point: – $P > 0.10 \Rightarrow$ non-significant evidence against H0 – $0.05 < P \leq 0.10 \Rightarrow$ marginally significant against H0 – $0.01 < P \leq 0.05 \Rightarrow$ significant evidence against H0 – $P \leq 0.01 \Rightarrow$ highly significant evidence against H0.

Two types of decision errors: Type I error = erroneous rejection of a true H0 Type II error = erroneous retention of a false H0

$\alpha \equiv$ probability of a Type I error

$\beta \equiv$ Probability of a Type II error

The traditional hypothesis testing paradigm considers only Type I errors. However, we should also consider Type II errors: $\beta \equiv$ probability of a Type II error, and $1 - b =$ "Power" $\equiv$ probability of avoiding a Type II error

Two forms of estimation

Point estimation ≡ single best estimate of parameter (e.g., x-bar is the point estimate ofμ)

Interval estimation ≡ surrounding the point estimate with a margin of error to create a range of values that seeks to capture the parameter; a confidence interval

Each sample derives a different point estimate and 95% confidence interval: 95% of the confidence intervals will capture the value of μ

- To create a 95% confidence interval for μ, surround each sample mean with a margin of error m that is equal to 2standard errors of the mean: $m \approx 2 \times SE = 2 \times (\sigma/\sqrt{n})$
- The 95% confidence interval for μ is now $x \pm m$
- Note that $\sigma/\sqrt{n}$ is the SE of the mean.

We rarely know population standard deviation $\sigma \Rightarrow$ instead, we calculate sample standard deviations s and use this as an estimate of σ. We then use s to calculate this estimated standard error of the mean:

- Using s instead of σ adds a source of uncertainty $\Rightarrow$ z procedures no longer apply $\Rightarrow$ use t procedures instead
- Standard Deviation (Error) of the Mean: The standard deviation of the sampling distribution of the mean has a special name: it is called the "standard error of the mean" (SE)
- The square root law says the SE is inversely proportional to the square root of the sample size:
- The sampling distribution of x-bar is Normal with mean μ and standard deviation (SE) = $\sigma / \sqrt{n}$ (when population Normal or n is large)

Analytical considerations for a trial

- Analyses by treatment group (adjusted for age, sex, race and center)
- Analyses with and ANCOVA (adjusts for age, sex, race and center, but further allows us to adjust for baseline values in the outcome)

If the pre-post change is the main outcome of interest, we can either

- Have the pre-post difference as the dependent variable
- Have the post measurement as the dependent variable, and adjust for baseline in the model
- Have the pre-post difference as the dependent variable, and further adjust for the baseline measurement in the model

Using differences versus ratios for a change in outcome as a dependent variable

- Count or differences in counts can give us the absolute change in an outcome variable
- Proportions are dependent of the total (count / total) so we need to adjust for the total count, to properly assess whether the ration changes over time because it depends on two variables (total may change from baseline while the ratio does not).
- We should not confuse numbers (counts, thresholds or ratios) that are clinically relevant with numbers (counts, thresholds or ratios) that are relevant for research purposes. An outcome may not be clinically significant, but may indicate significant results that will add to the literature.

Comparing Coefficients on an Additive VS Multiplicative Scale

When your dependent/independent variable is not normally distributed, you can either categorize the variable (at the cost of losing precision) or log transform the data (at the cost of interpreting the data on an additive/log vs multiplicative/normal scale). It is better to categorize your independent variables and log transform your dependent variables for ease of interpretation.

Example

In this page, we will discuss how to interpret a regression model when some variables in the model have been log transformed. The variables in the data set are writing, reading, and math scores (write, read and math), the log transformed writing (lgwrite) and log transformed math scores (lgmath) and female. For these examples, we have taken the natural log (ln). All the examples are done in Stata, but they can be easily generated in any statistical package. In the examples below, the variable write or its log transformed version will be used as the outcome variable. The examples are used for illustrative purposes and are not intended to make substantive sense.

Outcome variable is log transformed

Very often, a linear relationship is hypothesized between a log transformed outcome variable and a group of predictor variables. Written mathematically, the relationship follows the equation: $\log(y_i) = \beta 0 + \beta 1 * x1 + \ldots + \beta k * xk + e_i \ldots$ where y is the outcome variable and x1, .., xk are the predictor variables. In other words, we assume that $\log(y) - x'\beta$ is normally distributed, (or y is log-normal conditional on all the covariates.) Since this is just an ordinary least squares regression, we can easily interpret a regression coefficient, say $\beta 1$, as the expected change in log of y with respect to a one-unit increase in x1 holding all other variables at any fixed value, assuming that x1 enters the model only as a main effect. But what if we want to know what happens to the outcome variable y itself for a one-unit increase in x1? The natural way to do this is to interpret the exponentiated regression coefficients, $\exp(\beta)$, since exponentiation is the inverse of logarithm function.

- Let's start with the intercept-only model, log(write) = β0.

We can say that 3.95 is the unconditional expected mean of log of write. Therefore the exponentiated value is exp(3.948347) = 51.85. This is the geometric mean of write. The emphasis here is that it is the geometric mean instead of the arithmetic mean. OLS regression of the original variable y is used to to estimate the expected arithmetic mean and OLS regression of the log transformed outcome variable is to estimate the expected geometric mean of the original variable.

- Now let's move on to a model with a single binary predictor variable, log(write)= β0 + β1*female = 3.89 + .10*female

Now we can map the parameter estimates to the geometric means for the two groups. The intercept of 3.89 is the log of geometric mean of write when female = 0, i.e., for males. Therefore, the exponentiated value of it is the geometric mean for the male group: exp(3.892) = 49.01. What can we say about the coefficient for female? In the log scale, it is the difference in the expected geometric means of the log of write between the female students and male students. In the original scale of the variable write, it is the ratio of the geometric mean of write for female students over the geometric mean of write for male students, exp(.1032614) = 54.34383/49.01222 = 1.11. In terms of percent change, we can say that switching from male students to female students, we expect to see about 11% increase in the geometric mean of writing scores.

Arithmetic vs Geometric means

Definition of geometric mean: The geometric mean is the mean or average of a set of data measured on a logarithmic scale. The geometric mean is used when the logarithms of the observations are distributed normally (symmetrically) rather than the observations themselves.

The usual reason for choosing the geometric mean as a measure of location is to discount the influence of large observations. The geoMean is always less than the mean and that is sometimes an untold motivation for choosing it. Whether it is appropriate or not, the estimate is a random quantity and it can be characterized by a confidence interval. The little example below shows two methods for obtaining confidence intervals for both the arithmetic and geometric means. Here, the parametric method wrongly assumes that the data comes from a lognormal distribution)

- Proc GLM + Proc GENMOD: Means option gives you the unweighted means
- Proc GLM + Proc GENMOD: LSmeans option gives you the means accounting for covariates in the models
- Proc TTEST + Proc SURVEYMEANS: the dist = lognormal and allgeo options (respectively) give you geometric means for your linear outcome data

Tabular Methods Calculations

Pearson Chi Square

X^2 = [observed – expected]^2/expected

X^2 = [(ad-bc)^2)*T]/[n1n0m1m0]

Degrees of freedom = (row -1)(column – 1)

Logit Chi-square (Woolf)

X^2 used on the log scale because OR is not normaly distributed (i.e, 2 and ½ are at different distance for the null which is 1) but Ln(OR) is normally distributed (Ln(1) = 0 which is the null on the log scale)

Main effect

- Z = X-Xo/SE
- OR = e^B therefore, B = Ln(OR)
- Z = ln(OR) – Ln (1) / SE[ln(OR)], and **(Ln(1) = 0)**
- X^2 ~ Z^2 = [ln(OR)]^2/ SE[ln(OR)]^2
- X^2 = [ln(OR)]^2/ Var [ln(OR)] and **(SE^2 = Variance)**

95% CI

- Var [ln(OR)] = [1/a +1/b + 1/c + 1/d] So SE[ln(OR)] = √[1/a +1/b + 1/c + 1/d]
- 95% CI for the ln(OR): ln(OR) +/- 1.96*SE[ln(OR)] = ln(OR) +/- 1.96 * √[1/a +1/b + 1/c + 1/d]
- 95% CI for the OR: e^(ln(OR) +/- 1.96*SE[ln(OR)])

Using mantel-haenszel summary estimators

		Disease		
		Yes	No	
Exposure	Yes	a	b	m1
	No	c	d	m0
		n1	n0	Total

Let RRi be the stratum specific measure of association and let Estimator m-h be the pooled measure of association

Estimator = Σ(Wi*RRi)/ Σ (Wi)

- For Cumulative Incidence, Prevalence or Risk Ratio: RRmh = Σ (Wi*RRi)/ Σ (Wi) = Σ (c*m1/T)*[(a/m1)/(c/m0)]/ Σ (c*m1/T) = Σ (a*mo/T)/ Σ (c*m1/T)
- For Odds Ratio: ORmh = Σ (Wi*RRi)/ Σ (Wi) = Σ (b*c/T)*[(a/b)/(c/d)]/ Σ (b*c/T) = Σ [(a*d)/T]/ Σ [(b*c)/T]

How to test if two level of a dummy coded (or indicator) variable are different from one another:

The OR to compare B1 vs B2 (Ho: OR (B1 vs B2) =1 or Ho: B1 – B2 = 0) is

- OR (B1 vs B2) = Exp (B1 – B2) = exp (B1) / exp (B2)
- Z = B1 – B2 / Se (B1-B2)
- X^2 = (B1 – B2)^2 / Var (B1-B2)
- Since Se (B1 – B2) =/= Se(B1) – Se(B2)
- Var (B1-B2) = Var (B1) + Var (B2) – 2* Covar (B1,B2)

*** use the / COVB option in the model statement to get the covariance

- Use Se (B1) ^ 2 = Var (B1)
- Use Se (B2) ^ 2 = Var (B2)
- Get the Cov (B1, B2) from the /COVB option
- Find Var (B1 – B2) then Se (B1 – B2)
- Find Z then X ^2 and p-value

M[illegible]-Haenszel method for adjustment

Review

[illegible]alculated adjusted estimate is a weighted average (a pooling) of the stratum specific estimates

General form

$$\text{ESTIMATE}_{MH} = \frac{\sum(\text{Weight}_i \times \text{Estimate}_i)}{\sum(\text{Weight}_i)}$$

For relative risks (incidence proportion ratio or prevalence ratio)

$$\text{PREVALENCE RATIO}_{MH} = \frac{\sum\left(\frac{c_i n_{1i}}{N_i}\right)\left(\frac{a_i / n_{1i}}{c_i / n_{0i}}\right)}{\sum\left(\frac{c_i n_{1i}}{N_i}\right)} = \frac{\sum\left(\frac{a_i n_{0i}}{N_i}\right)}{\sum\left(\frac{c_i n_{1i}}{N_i}\right)}$$

	D+	D-	
E+	a_i	b_i	n_{1i}
E-	c_i	d_i	n_{0i}
			N_i

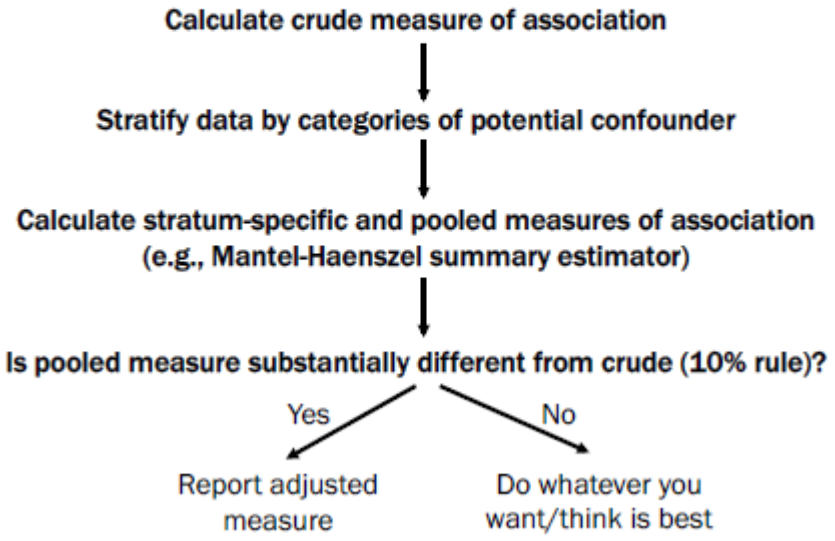
10

Unlike confounding, EM is something we want to describe, report, & understand…

- Interaction can reflect real mechanistic features in the causal framework
- Or it can be a statistical quality of a situation
  - This doesn't mean there aren't PH implications to the interaction!
- (Unlike confounding which just comes from an imbalance of a 3rd variable which is a cause of the outcome and happens to have an unbalanced distribution across exposure levels)
- (Also unlike bias which comes from mistakes you made in study design or analysis)

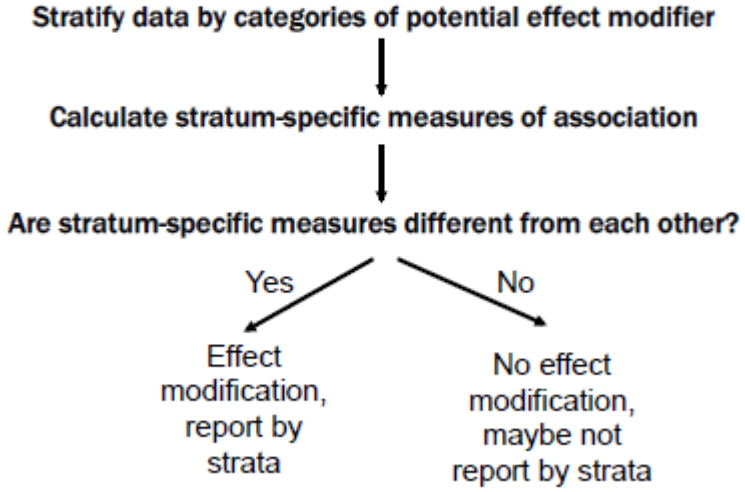
15

Path for assessing and reporting confouding using tablular methods



17

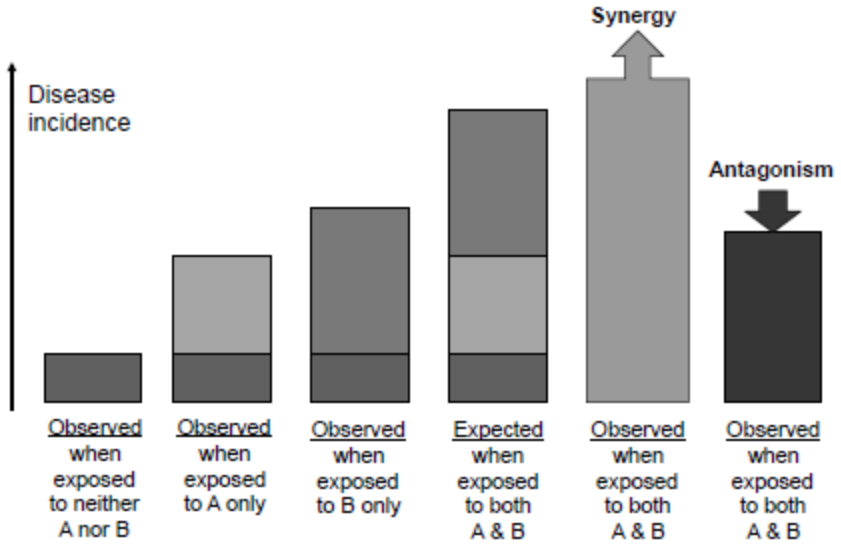
Path for assessing and reporting interaction using tablular methods



18

Synergy and antagonism

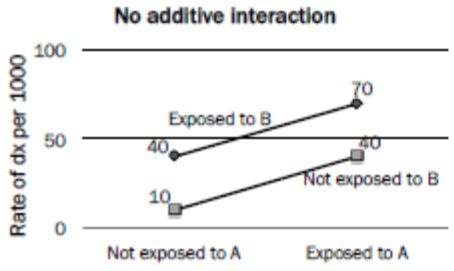
Synergy = more than the combo, Antagonism = less than the combo. But how do you define "combo?"

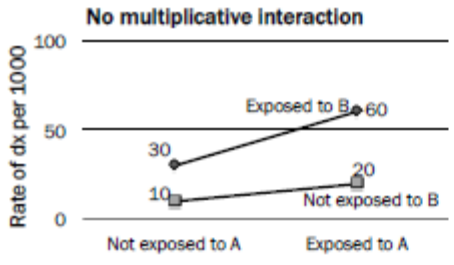


25

Is there interaction?

Let's check for additive and multiplicative interactions





Additive scale

$$S = \frac{RR_{11} - 1}{(RR_{10} - 1) + (RR_{01} - 1)}$$

Multiplicative scale

$$\frac{RR_{11}}{(RR_{10} * RR_{01})}$$

$RR_{10} = R_{10}/R_{00}$
$RR_{01} = R_{01}/R_{00}$
$RR_{11} = R_{11}/R_{00}$

- 1 = no interaction
- > 1 = synergistic or positive
- < 1 = antagonistic or negative

29

Analysis of Variance

Problem of Multiple Comparisons

Family-wise error rate = probability of at least one false rejection of H0. Assuming three null hypotheses are true:

- At $\alpha = 0.05$, the Pr(retain all three H0s) = $(1-0.05)3 = 0.857$.
- Therefore, Pr(reject at least one) = $1-0.847 = 0.143 \Rightarrow$ this is the family-wise error rate.
- The family-wise error rate is much greater than intended. This is "The Problem of Multiple Comparisons"

One-way Analysis Of Variance (ANOVA)

- Categorical explanatory variable
- Quantitative response variable
- Test group means for a significant difference

Method: compare variability between groups to variability within groups (F statistic)

Variability between Groups

- Variability of group means around the grand mean → provides a "signal" of group difference
- Based on a statistic called the Mean Square Between (MSB)
- Mean Square Between = Sum of Squares Between [Groups] / Degrees of Freedom Between

Variability within Groups

- Variability of data points within groups → quantifies random "noise"
- Based on a statistic called the Mean Square Within (MSW)
- Mean Square Within = Sum of Squares within [Groups] / Degrees of Freedom within

***F* statistic** is the ratio of the *MSB* to *MSW (Aka the signal to noise ratio)*

- ANOVA for two groups is equivalent to the equal variance (pooled) *t* test . ANOVA *H*a says "at least one population mean differs" but does not delineate which differ. Post hoc comparisons are pursued after rejection of the ANOVA *H*0 to delineate differences
- Bonferroni Procedure: The Bonferroni procedure is instituted by multiplying the P-value from the LSD procedure by the number of post hoc comparisons "c".
- The Kruskal-Wallis test is the non-parametric analogue of one-way ANOVA. It does not require Normality or Equal Variance conditions for inference. It is based on rank transformed data and seeing if the mean ranks in groups differ significantly.

Extension of an ANOVA analysis

Look at the effect of ABO blood groups on inflammation markers after controlling for other cancer risk factors among pancreatic cancer and colorectal cancer patients only.

Will use a permutation based ANCOVA approach. This will allow us to determine the effect of one independent variable (ABO blood group) on a dependent, continuous or scale variable (here the NLR, FVIII, VWF ect..) in people with pancreatic or colorectal cancer, after controlling for other covariates. (Katz, Parish and Williams, The Effect of Race/Ethnicity on the Age of Colon Cancer Diagnosis, 2013)

Multivariate analysis of covariance (MANCOVA) is a statistical technique that is the extension of analysis of covariance (ANCOVA). Basically, it is the multivariate analysis of variance (MANOVA) with a covariate(s).). In MANCOVA, we assess for statistical differences on multiple continuous dependent variables by an independent grouping variable, while controlling for a third variable called the covariate; multiple covariates can be used, depending on the sample size. Covariates are added so that it can reduce error terms and so that the analysis eliminates the covariates' effect on the relationship between the independent grouping variable and the continuous dependent variables.

Questions answered by MANOVA can be:

- Do the various school assessments vary by grade level after controlling for gender?
- Do the rates of graduation among certain state universities differ by degree type after controlling for tuition costs?
- Which diseases are better treated, if at all, by either X drug or Y drug after controlling for length of disease and participant age?

Sample Size Estimation

Estimation of sample size is one of the most common questions posed of statisticians, yet it is almost entirely a non-statistical exercise. The following are important points to remember:

1. Sample size should be specified in advance of the study; the number is usually determined by the primary objective of the study;

2. Sample size estimation required knowledge of the subject area and collaboration;

3. Sample size should account for non-compliance and withdrawal from the study because the primary analysis should be intent-to-treat..

4. Parameters on which sample size is based (apart from treatment differences) should be evaluated as part of interim monitoring of the study by the investigators; and

5. In the long run, it pays to be conservative in estimating sample size.

α = Type I error or significance level - probability that the trial will find two treatments that are equally effective "significantly" different from one another.

β = Type II error - probability of failing to reject the null hypothesis when there is a difference between two treatments of the size specified as clinically relevant (Delta); power = 1 - β.

"Variability" refers to estimated variance of outcome measure. For morbidity/mortality outcomes variability is a function of how many events are expected to occur.

"Delta" is the size of the difference between the experimental and control groups considered important -- a difference it is important not to miss. To describe some of the thinking which goes into sample size estimation, estimation of sample size for a continuous response variable such as blood pressure will be illustrated for parallel groups and crossover designs (Section I). In Section II issues of sample size estimation in long term trials with morbidity and mortality outcomes will be considered.

$Z_{1-\alpha/2}$ = the value of the Normal distribution which cuts off an upper tail probability of $\alpha/2$. For $\alpha = 0.05$, $Z_{1-\alpha/2} = 1.96$.

$Z_{1-\beta}$ = the value of the Normal distribution which cuts off an upper tail probability for β. For $\beta = 0.2$, $Z_{1-\beta} = 0.84$.

Rewrite breaking up the total error into that due to:

σ_S^2 - patient variability (between-patient)

σ_e^2 - measurement error and temporal variation (within-patient)

$$\text{(n for each group)} = \frac{2(\sigma^2_S + \sigma^2_e)\,(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\Delta^2}$$

Calculating Power

- Null hypothesis testing proposes a theory presumed to be true (e.g., treatment has no effect) and then attempts to reject that theory.
- There are only two outcomes: reject the null or fail to reject the null.
    - Type I Error (a) is the error of rejecting the null hypothesis when it is actually true. (False positive)
    - Type II Error (b) is the error of failing to reject the null hypothesis when the alternative hypothesis is true. (False negative)
    - Statistical Power is the probability that the null hypothesis will be rejected when the null hypothesis is false. (1 – b)
- When designing a study, multiple factors to consider
    - α
    - Power (1 – β)
    - Sample Size
    - Expected Effect Size
- Often constrain a=.05, power = .8 (or .9), and determine the necessary sample size to detect the expected effect size.
- Alternately, can use a, power, and sample size to determine how big an effect can be detected (common with secondary data analysis)

Formula to Calculate Sample Size for Two Sample Mean Test

$$m = (2 * \sigma^2_Y * (t_{crit\alpha} + t_{crit\beta})^2) / \Delta^2$$

where
m = N per group
σ^2_Y = variance of Y
$t_{crit\alpha}$ = 1.96 (for α=.05 two-tailed)
$t_{crit\beta}$ = .85 (for β=.20)
Δ^2 = Square of the expected effect/difference

You can determine the variance of Y and the expected effect by using the litteraure. You can back-calculate the variance if you have a 95% CI

- UB = $\mu + 1.96 * \sigma$
- LB = $\mu - 1.96 * \sigma$
- Therefore σ = [UB – μ]/ 1.96 or [LB+ μ]/1.96

Randomization

Randomization helps us control for measured and unmeasured confounders by making their distribution among our exposed and unexposed groups similar.

- Simple randomization allows individual allocation to treatment or control group using simple mechanisms such as a coin flip or a randomization table
- Block randomization creates small groups (blocks) in which there are equal number or people in both treatment and control blocks. Balances number of people in each arm during recruitment
- Stratification followed by block randomization balances the stratification factor (age, severity of the disease) across treatment arms

Sensitivity

Sensitivity = a / a+c = proportion of people with the classification who are correctly identified

Specificity = b / b+d = proportion of people without the classification who are correctly identified

Define false positives and false negatives as well as sensitivity and specificity. Describe the mathematic relationship between sensitivity and false negative, and specificity and false positive.

Healthy is not the same as a "normal" test – Healthy with normal test (true negative) – Healthy with abnormal test (false positive)

Disease is not the same as "abnormal" test – Sick with abnormal test (true positive) – Sick with normal test (false negative)

- Sensitivity is probability that a test correctly classifies as positive individuals with disease (Tp/Tp+Fn)
- Specificity is the probability that a test correctly classifies individuals without disease as negative (Tn/Tn+Fp)

Know the criteria that are used to assess the value of a screening test.

- (1-sensitivity) = False negative rate
- (1-specificity) = False positive rate

Person Time Calculation

Definition: A measurement combining the number of persons and their time contribution in a study. This measure is most often used as denominator in incidence rates. It is the sum of individual units of time that the persons in the study population have been exposed or at risk to the conditions of interest. The most frequently used person-time is person-years.

Person time can be calculated using the information below:

- If a person develops the disease on day 2, they contribute 1.5 person-days during which they were at risk for developing disease

- If a person is at risk for 30 days and does not contract the disease, they contribute 30 person-days at risk
- Combined, these two people contributed 31.5 person days at risk

Directly calculating person-time is tedious at best and often impossible. We can estimate person-time using the following formula:

[(Number of people at risk at the beginning of the time interval +

Number of people at risk at the end of the time interval) / 2] x (Number of time units in the time interval)

Example 1: A population at risk is composed of 100 senators. Twenty-five senators develop symptoms consistent with inhalation anthrax disease and are confirmed by laboratory testing to have been infected with Bacillus anthracis. If 12 senators developed anthrax in September and 13 developed anthrax in October, what is the incidence rate of anthrax for those two months? In this case:

- Numerator is the 25 new cases
- Denominator (person-time at risk) could be calculated by: [(100 Senators at risk at the beginning of Sept. + 75 Senators at risk at the end of Oct.) / 2] x 2 months = [(175 / 2) x 2] months = 175 person-months of risk
- Note: Since 25 Senators got anthrax in September and October, there are 75 Senators remaining at risk at the end of October.

The incidence rate would then be: (25 new cases) / (175 person-months of risk) = 14% of the senators are getting anthrax each month.

Example 2: Assume 5 participants were followed for 5 years, and 3 developed prostate cancer. The time contributed by each subject is as follows:

- Subject A: 2.5 years
- Subject B: 5 years
- Subject C: 1.5 years
- Subject D: 5 years
- Subject E: 0.5 years

Total person-years in the study: (2.5+5+1.5+5+0.5) = 14.5 person-years

14.5 p-y is the denominator in the rate of prostate cancer. The rate is 3/(14.5 p-y), or 0.207 cases per p-y. By multiplying both the numerator and denominator by 1000 the rate becomes 207 cases per 1000 p-y.

Person Time Terminology

Rate: the number of new cases of disease during a period of time divided by the person-time-at-risk

Person-time: estimate of the actual time-at-risk in years, months, or days that all persons contributed to a study

Standardization VS IPW

Age	Stroke	N	Dementia	N	Dementia
65-74	No	15000	750	35000	3500
65-74	Yes	4000	1400	6000	1800
75+	No	18000	3600	12000	3000
75+	Yes	8000	4400	2000	1000
Total		45000	10150	55000	9300

A = Stroke, L = Age, Y = Dementia

Question 1

Standardization: $\sum Pr[Y=1, A=a \& L=l] * Pr[L=l]$ (there are 4 strata)

Standardization weight per strata = (n in treatment group of the strata + n in control group of the strata) / N in all study

For example, in the 65-74 strata, it will be (15000 + 35000)/(45000 + 55000) = 0.5

Age	Stroke	N	Dementia	N	Dementia	Standardization Weight	Risk Difference	$\sum Pr[Y=1, A=a \& L=l] * Pr[L=l]$
65-74	No	15000	750	35000	3500	0.5	-0.05	-0.025
65-74	Yes	4000	1400	6000	1800	0.1	0.05	0.005
75+	No	18000	3600	12000	3000	0.3	-0.05	-0.015
75+	Yes	8000	4400	2000	1000	0.1	0.05	0.005
Total		45000	10150	55000	9300			-0.03

The risk difference in dementia between the brain boost group and the placebo group is -0.03. The Brain boost group will have 3 fewer cases per 100 cases of dementia, compared to the placebo group.

Question 2

Inverse Probability Weighting = P(A=a given L=l) for cases and controls separately to create pseudo population. within a strata, it is n1/(n1+n2)
For example, in the 65-74 strata, it will be (15000)/ (15000 + 35000) = 0.3 weight for treatment group
For example, in the 65-74 strata, it will be (35000)/(15000 + 35000) = 0.7 weight for the control group
Then take the inverse of the weights (hence the inverse probability weighting name) to multiply the number of cases and participants in each strata (creates a pseudo population) which you can then use to calculate your measures of association regularly.

Age	Stroke	N	Dementia	N	Dementia	Weight for treatment	Weight for Placebo	Inverse Weight for Treatment	Inverse Weight for Placebo	Pseudo N for treatment group	Dementia in Treatment Group	Pseudo N for placebo group	Dementia in placebo Group
65-74	No	15000	750	35000	3500	0.3	0.7	3.33	1.42	50000	2500	50000	5000
65-74	Yes	4000	1400	6000	1800	0.4	0.6	2.5	1.67	10000	3500	10000	3000
75+	No	18000	3600	12000	3000	0.6	0.4	1.67	2.5	30000	6000	30000	7500
75+	Yes	8000	4400	2000	1000	0.8	0.2	1.25	5	10000	5500	10000	5000
Total		45000	10150	55000	9300					100000	17500	100000	20500

The risk difference in dementia between the brain boost group and the placebo group is -0.03. The Brain boost group will have 3 fewer cases per 100 cases of dementia, compared to the placebo group.

Question 3

Standardization: $\sum Pr[Y=1, A=a \& L=l] * Pr[L=l]$ (there are 4 strata)
Standardization weight per strata = n for treatment group in the strata + n for control group in the strata / N in all study

Age	Stroke	N	Dementia	N	Dementia	Standardization Weight	Risk Difference	$\sum Pr[Y=1, A=a \& L=l] * Pr[L=l]$
65-74	No	15000	750	35000	3500	0.5	-0.05	-0.025
65-74	Yes	4000	1400	6000	1800	0.1	0.05	0.005
75+	No	18000	3600	12000	3000	0.3	-0.05	-0.015
75+	Yes	8000	4400	2000	1000	0.1	0.05	0.005
Total		45000	10150	55000	9300			

For the 65 to 74 and No stroke strata and the over 75 and no stroke strata, the risk difference between the brainboost group and the placebo group was -0.05. The Brain boost group would have 5 fewer cases per 100 cases of dementia, compared to the placebo group. However for the 65 to 74 and Stroke strata and the over 75 and Stroke strata, the risk difference between the brainboost group and the placebo group was 0.05. The Brain boost group would have 5 more cases per 100 cases of dementia, compared to the placebo group.

Question 4

The effects estimated in question 1 is the absolute risk difference in dementia averaged over all four strata in our study, using the population as a standard. In the presence of exchangeability this measure can be used to estimate the (counterfactual) risk that would have been observed had everyone in the population taken brainboost. The effects estimated in question 2 represent the absolute risk difference in by creating a pseudo population in which we observed the outcomes if all participants had taken brainboost, and then we saw their outcomes if they had all taken placebo. Both standardization in question 1 and IPW in question 2 allowed us to simulate what would have been observed if we did not have different probability of being assigned in to strata based on Age and Stroke history in this study, and thus compute the average causal effects (assuming positivity, consistency and exchangeability) and gave us the same answer. In question three, we evaluated the absolute risk difference in each individual strata, taking into account that the vector composed of age and stroke status variables affected the probability of treatment.

Question 5

Based on the previous results, I would only recommend Brainboost among those 65 years and older if they have no previous history of stroke. Even though we observed an overall benefit in the absolute risk in dementia in our population, we consistently observed benefits (-0.05 *50000 + -0.05 * 30000 = 4000 fewer cases of dementia thanks to brain boost) among those without a stroke history, but no benefits for those with a history of a stroke (the positive risk difference implies that brain boost had no protective effect in the 65-74 and Stroke or the 75+ and Stroke groups). The only caveat would be that the observed risk differences would have to be collapsible over the age group categories so that they would remain the same if we were to look at Stroke history alone.

Measures of Incidence

Cumulative Incidence = # of new cases / # at risk = a+c / a+b+c+d

Incidence Rate = # of new cases / person-time at risk = a / PY1

Prevalence = # new cases + existing cases / Defined population

Mathematically, Prevalence = Incidence Rate * Duration

When the disease in rare (a and c are small) so:

CIR = (a/a+b) / (c/c+d) ~ (a/b) / (c/d) = OR

Measures of association

Absolute measures of association are differences, and are valuable for public health decision making and policy

- Absolute Risk = cumulative incidence in exposed - cumulative incidence in unexposed = CID
- Absolute Rate = incidence rate in exposed - incidence rate in unexposed = IRD

Relative measures of association are ratios, and are indicators of the strength of an association (etiological epi, where 1 exposure > 1 disease). Incidence Rates are better for dynamic populations, whereas cumulative incidence is better for fixed populations and is easier to interpret

- Relative Risk = cumulative incidence in exposed / cumulative incidence in unexposed = CIR
- Relative Rate = incidence rate in exposed / incidence rate in unexposed = IRR

Excess Risk: CI(in exposed) – CI(in unexposed) / CI(in unexposed) = (a/a+b) – (c/c+d) / (c+c+d) = Percentage in risk from baseline i.e. unexposed population as a result of the exposure

Measures of impact

Measures of impact are important to predict the impact of removing a particular exposure on the risk of developing an outcome.

- Rate Difference = CI1 – CI0 = CID. Therefore (CI1 – CI0)*(a+b) = CID*n(exposed) = number of new cases among the exposed due to the exposure
- Population Rate Difference: proportion exposed = (a+b) / (a+b+c+d) = prevalence of an exposure. PRD = (CI1 – CI0)*(proportion exposed) = CID*p1. Therefore PRD*N = CID*P1*N = number of new cases in the whole population that were due to the exposure
- Attributable proportion among the exposed = (CI1 – CI0)/CI1 = CIR – 1 / CIR. APe is the percentage of cases due to the exposure, among the exposed
- Attributable proportion among the total population = (CI - CI0)/CI = p(CIR – 1) / 1+p(CIR-1). APt is the percentage of cases due to the exposure, among the whole population

Association Tests

	Categorical Exposure	Continuous Exposure
Categorical Outcome	Chi-Square Test	Logistic Regression Multinomial Regression
Continuous Outcome	T-test ANOVA	Regression

Hypothesis Testing

Truth		Test	
		Non-Significant	Significant
	Ho	True Negative	Type 1 Error (False Positive)
	Ha	Type 2 Error (False Negative)	True Positive

General Linear Models & Distribution Families in Statistics

The Generalized Linear Model

The link function g is usually chosen so that $g(\mu_i)$ can take on values in $(-\infty, +\infty)$:

- **Bernoulli (binary) data:** $g(\mu) = logit(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
- **Poisson (count) data:** $g(\mu) = \log(\mu)$
- **Normal (continuous) data:** $g(\mu) = \mu$
- **Encountered less frequently:**
  - $g(\mu) = 1/\mu$ (inverse; positive continuous data)
  - $g(\mu) = \sqrt{\mu}$ (square root; count data)
  - $g(\mu) = \Phi(\mu)$ (probit; Bernoulli/Binomial data)

The Generalized Linear Model

Tackling the variance problem

Specify a **variance function** relating the variance of the outcome to the mean:

$$Var(Y_i \mid \mathbf{x}_i) = f_V(\mu_i)$$

Examples:

- **Bernoulli (binary) data:** $f_V(\mu) = \mu(1-\mu)$
- **Poisson (count) data:** $f_V(\mu) = \mu$
- **Normal (continuous) data:** $f_V(\mu) = 1$

Common Covariance Patterns

- Compound Symmetry
  - Common variance at each time
  - Common covariance for each $cov_{i,j}$
  - 2 parameters: σ^2 and σ_1

$$\begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{bmatrix}$$

Common Covariance Patterns

- Toeplitz
  - Common variance at each time
  - Unique covariance for each $cov_{i,j}$ 'band'
  - t parameters: σ^2 and t-1 covariances

$$\begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}$$

Common Covariance Patterns

- First order Autoregressive
  - Common variance at each time
  - Covariance for each $cov_{i,j}$ decays exponentially with distance apart in time
  - 2 parameters: σ^2 and ρ

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Common Covariance Patterns

- Unstructured
  - Unique variance at each time
  - Unique covariance for each $cov_{i,j}$ pair
  - t(t+1)/2 parameters
  - Quickly becomes a large number !

$$\begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

General Data Analysis Review

Distribution Families in Statistics

Data Type	Distribution	Model	Mean Function	Variance Function	Canonical Link Function	Measure of Association
Continuous	Gaussian	Linear Model	g(u) = u	f(u) = 1	Identity Link	Risk Difference
Positive Continuous	Exponential	Gamma Regression	g(u)=1/u	f(u) = u^2	Inverse Link	Inverse Risk Difference
Binary (Bernoulli)	Bernoulli	Logistic Regression	g(u) = Log (u/1-u)	f(u) = u(1-u)	Logit Link	Odds Ratio
Binary (Discrete values)	Poisson	Poisson Regression	g(u) = Log(u)	f(u) = u	log link	Relative Risk

There are typically three types of statistical models that we commonly use in epidemiologic data analyses

	Distribution	Link	Mean Function	Variance Function	Estimate
Gaussian	Normal	Identity	g(u) = u	var(u) = 1	Risk Difference
Logistic	Binary	Logit	g(u) = u/(1-u)	var(u) = u(1-u)	Odds Ratio
Poisson	Binary	Log	g(u) = log(u)	var(u) = u	Relative Risk

Distribution options in SAS

DIST=	Distribution	Default Link Function
BINOMIAL | BIN | B	binomial	logit
GAMMA | GAM | G	gamma	inverse (power(1))
GEOMETRIC | GEOM	geometric	log
IGAUSSIAN | IG	inverse Gaussian	inverse squared (power(2))
MULTINOMIAL | MULT	multinomial	cumulative logit
NEGBIN | NB	negative binomial	log
NORMAL | NOR | N	normal	identity
POISSON | POI | P	Poisson	log

Statistical Software in Use

- SAS: Data Cleaning + Data Analysis
- STATA: Secondary Analyses (Interactions/Correlated & Clustered Data/Splines/G-Methods)
- R: Genetic SNPs analyses + Microbiomic Taxa Analyses

Before the analysis

- Do your literature review to see how the outcome of interests / similar analyses were reported in the literature
- Make Blank tables for your analysis in an excel spreadsheet (Table 1 for demographics, Table 2 for main model, Table 3 for secondary/stratified analyses, and figures when appropriate i.e. interaction, clustered data) based on published papers in the litterature

Analysis Step 1: EDA

- Put the data in wide format (Each person has one row)
- Run descriptive statistics (mean, variance, histogram, boxplots) on the outcome
- Run descriptive statistics of your exposure variables (for your Table 1, make sure that covariates are equally distributed across Txt arms or case/control status) and transform them if necessary
- Observe the correlation in the outcomes (pearson correlation coefficients/scatterplots of repeated measures/ Calculate the ICC and Covariance Matrices)
- Cross tabulate or run simple t-tests for your outcome vs your exposures

Analysis Step 2: Modeling

- Put the data in long format (each correlated observation has its own row)
- Based on your EDA and outcome type (continuous, binary, categorical or count) pick your regression model
- Run crude model between outcome and main exposure -> Run adjusted models, and properly specify correlation if needed
- Look at stratified analyses (Breslow day for stratum specific estimates) -> Look at pooled estimates interactions -> Look for potential confounding
- Make figures and graphs where appropriate

Analysis step 3: Model Troubleshooting

- If your model does not converge, check your outcome variables to make sure you are using the appropriate link function for the data (linear, logistic, poisson, Hazard, cox, ect…)
- The second step is to check for complete separation in your data (on the continuous scale, when your predictor perfectly separate your outcome values: more common with binomial outcomes, try to make bigger categories in your predictors to remediate that)
- The last step is to change the convergence criterion: Least square means > Maximum Likelihood > Restricted Maximum Likelihood

After the analysis

- Give a detailed report on the methods used in the methods section, and make sure you explicitly state the assumptions, limitations and biases in your discussion section

Statistical Testing Considerations in Molecular Biology and Bioinformatics

In the field of genomics (and more generally in bioinformatics), the modern usage is to define fold change in terms of ratios and not by the alternative definition.

However, log-ratios are often used for analysis and visualization of fold changes. The logarithm to base 2 is most commonly used as it is easy to interpret, e.g. a doubling in the original scaling is equal to a $\log_2$ fold change of 1, a quadrupling is equal to a $\log_2$ fold change of 2 and so on. Conversely, the measure is symmetric when the change decreases by an equivalent amount e.g. a halving is equal to a $\log_2$ fold change of -1, a quartering is equal to a $\log_2$fold change of -2 and so on. This leads to more aesthetically pleasing plots as exponential changes are displayed as linear and so the dynamic range is increased. For example, on a plot axis showing $\log_2$ fold changes, an 8-fold increase will be displayed at an axis value of 3 (since 2^3 = 8). However, there is no mathematical reason to only use logarithm to base 2, and due to many discrepancies in describing the $\log_2$ fold changes in gene/protein expression, a new term "loget" has been proposed.

Fold change is a measure describing how much a quantity changes between an original and a subsequent measurement. It is defined as the ratio between the two quantities; for quantities A and B, then the fold change of B with respect to A is B/A. Fold change is often used when analysing multiple measurements of a biological system taken at different times as the change described by the ratio between the time points is easier to interpret than the difference.

Fold change is so-called as it is common to describe an increase of multiple X as an "X-fold increase". As such, several dictionaries, including the Oxford English Dictionary and Merriam-Webster Dictionary, as well as Collins's Dictionary of Mathematics, define "-fold" to mean "times," as in "2-fold" = "2 times" = "double." Likely because of this definition, many scientists use not only "fold" but also "fold change" to be synonymous with "times", as in "3-fold larger" = "3 times larger.". More ambiguous is fold decrease, where for instance a decrease of 50% between two measurements would generally be referred to a "half-fold change" rather than a "2-fold decrease".

Fold change is often used in analysis of gene expression data from microarray and RNA-Seq experiments for measuring change in the expression level of a gene. A disadvantage and serious risk of using fold change in this setting is that it is biased and may misclassify differentially expressed genes with large differences (B-A) but small ratios (B/A), leading to poor identification of changes at high expression levels. Furthermore, when the denominator is close to zero, the ratio is not stable and the fold change value can be disproportionately affected by measurement noise.

With these microarrays you have measured the expression of the miRNAs in two conditions. The ratio of theses expression values ("treatment" condition vs. "reference" condition) if called the "fold-change" (FC). It's logarithm is called the log fold-change, abbreviated logFC. The logFC is the more attractive measure for differential expression than the FC, because

- its error distribution is symmetric (not skewed)
- "zero" (0) means "no change", positive values indicate up- and negative value indicate down-regulation

- similar absolute values (e.g. -2 and +2) can be seen as effects of similar biological relevance (only in different directions)
- similar differences between values can be seen as differential effects of similar biological relevance

The logarithm in the logFC is typically calculated for the base 2. That means one unit of the logFCs translates to a two-fold change in expression. The FCs can be calculated from the logFCs as FC = 2^logFC.

Sometimes, not only the regulation (differential expression) of the miRNA is interesting but also the general level at which it is expressed. This is giben by the geometic average of the expressions under both conditions, what is the square-root of the product of the two expression values. On the logarithmic scale this translates simply to the average. The result is shown in the colum AveExp (for Average (log-)Expression). The higher the value, the higher is the general expression (abudance, concentration) of the miRNA.

The logFC was determined as the mean of several samples. The result scatters around an (unknown) expected value and is used as an estimate for this unknown value. The precision (or variability) of this estimate is usually indicated either by the standard error (SE) or (better) by a confidence interval. These values are missing in the table you got.

The estimated logFC can be compared to a hypothesized value using a t-test. The null hypothesis is typically that the gene is not regulated, so the expected value is zero. For this hypothesis a test statistic with known distribution can be calculated. This is the t-value which is calculated as t = logFC/SE. So if desired you can retrieve the missing SE simply by dividing the logFC by t. For the t-value a correspoding p-values can be calculated. The p-value is the probability, under the null hypothesis, to get more extreme t-values than the one calculated. P-values close to zero indicate that the obtained t-value is unlikely if the miRNA was not regulated. By some strange inverse (and wrong) logic most people conclude that this would demonstrate that the miRNA is regulated (differentially expressed). The story behind the interpretation of P-values is long and complicated and should not be discussed here. You can find several enlightning threads about this topic in ReseachGate.

The B-value is another statistic about the regulation. It is sometimes called "log odds ratio" and gives the logarithm of the ratio of the odds for regulation to the odds against regulation. Similar to the logFC, B-values of 0 indicate 50:50 odds (maximally undecided), positive values indicate that the data favours up regulation, negative values indicates that the data favours down regulation. The B-value is based on Bayesian theorem (therefore possibly the abbreviation "B") and the prior expectation that some particular fraction of the miRNAs is regulated. This proportion should also be given by these bioinformatics people (actually you should have told them which value to use, because this depends on your expert judgement and not on statistics).

No, there is no general objective justification for any particular log-fold change threshold. Mathematically speaking, it is possible to reject the null hypothesis at any non-zero log-fold change if the variability is low enough. One could argue that small log-fold changes are not biologically relevant, but the exact definition of "small" is open to interpretation. Larger log-fold changes are also more robustly detected across technologies (e.g., RNA-seq and qPCR), though

selecting a threshold on this basis would depend on the sensitivities of the technologies involved. Somewhere between 1.1 to 1.5 is a common choice for a "sensible" threshold.

But all this is getting away from the main point, which is the detection of DE genes. If you want to do this in a statistically rigorous manner, use the BH-adjusted p-values to control the false discovery rate. This ensures that the expected proportion of false positives in your set of significant DE genes is below a certain threshold (usually 5%). Now, you might say that this approach also involves the selection of an arbitrary threshold. However, with this approach, at least the choice of threshold is directly related to the probability of whether the genes are truly DE or not. A log-fold change threshold doesn't tell you much about the error rate, as it doesn't account for the variability of the expression values.

Finally, if you do need a log-fold change threshold, the treat function should be used, and DE genes selected on the basis of the adjusted p-values. This ensures that the FDR is controlled while only considering genes with log-fold changes above a minimum value.

Controlling for multiple testing

For 1 test, we typically set the alpha value at 0.05

This means that if we run 100 test, at least 5 of them will be statistically significant

A p-value is an area in the tail of a distribution that tells you to odds of a result happening by chance.

Bonferroni test

Divide the alpha value by the number of test being done: more stringent, and it controls the probability of having one or more false positive

False discovery adjustment

Guarantees the expected number of false positive, independent of the number of test

So if your FDR is at 0.2, then 20% of all the q-values you call as hits will be false positives

A Q-value is a p-value that has been adjusted for the False Discovery Rate(FDR). The False Discovery Rate is the proportion of false positives you can expect to get from a test. A p-value gives you the probability of a false positive on a single test; If you're running hundreds or thousands of tests from small samples (which are common in fields like genomics), you should use q-values.

Why are Q-Values Necessary?

Usually, you decide ahead of time the level of false positives you're willing to accept: under 5% is the norm. This means that you run the risk of getting a false statistically significant result 5% of the time. You can't escape this fact when you're running tests: false positives (p-values) are a fact of life and are unavoidable. While 5% might be an acceptable false positive rate for running one test, it becomes completely unacceptable if you run thousands of tests on the same small data set. Here's why:

Imagine you're planning scratch off lottery, and you have a 5% chance of getting a winning ticket. One ticket gives you a 5% chance, but if you buy enough tickets, probability tells us that you'll eventually get a winner (buying 1,000 lottery tickets should do the trick and will in fact give you, on average, 50 winning tickets). The same is true for lab tests.

- The first test on your data, you have a 5% chance of a false positive.
- The second test on your data, you have another 5% chance of a false positive.
- The thousandth test on your data, you have had a 5% chance of a false positive a thousand times.

Essentially, you'll get a false positive — a false "significant" result — if you run enough tests. In fact, at a 5% FDR, you'll get 5 false results for every 100 tests you run, or 50 for every thousand. That's pretty high. This is called the multiple testing problem.

The False Discovery Rate approach to p-values assigns an adjusted p-value for each test. This is the "q-value." A p-value of 5% means that 5% of all tests will result in false positives. A q-value of 5% means that 5% of significant results will result in false positives. Q-values usually result in much smaller numbers of false positives, although this isn't always the case.

To put this another way, p-values tell you the percentage of false positives to expect and take into account the number of tests being run. For example, if you run 1600 tests, you would expect to see about 80 false positives. The q-value doesn't take into account all the tests; they only take into account the tests that are below a threshold that you choose (i.e. tests reporting a q-value of 5% or less).

Note: The Q-value is not the same as the "Q" you sometimes see in statistics. Q on its own (as opposed to a Q-value) refers to elements in a set that don't have a particular attribute. For example, let's say you had 100 people and 57 of them like pizza. The proportion of people who like pizza is P=0.57. Therefore, Q = 0.43 (which is just 1 – P).

What are p-values?

The object of differential 2D expression analysis is to find those spots which show expression difference between groups, thereby signifying that they may be involved in some biological process of interest to the researcher. Due to chance, there will always be some difference in expression between groups. However, it is the size of this difference in comparison to the variance (i.e. the range over which expression values fall) that will tell us if this expression difference is significant or not. Thus, if the difference is large but the variance is also large, then the difference may not be significant. On the other hand, a small difference coupled with a very small variance could be significant. We use the one way Anova test (equivalent t-test for two groups) to formalise this calculation. The tests return a p-value that takes into account the mean difference and the variance and also the sample size. The p-value is a measure of how likely you are to get this spot data if no real difference existed. Therefore, a small p-value indicates that there is a small chance of getting this data if no real difference existed and therefore you decide that the difference in group expression data is significant. By small we usually mean 0.05.

What are q-values, and why are they important?

False positives

A positive is a significant result, i.e. the p-value is less than your cut off value, normally 0.05. A false positive is when you get a significant difference when, in reality, none exists. As I mentioned above, the p-value is the chance that this data could occur given no difference actually exists. So, choosing a cut off of 0.05 means there is a 5% chance that we make the wrong decision.

The multiple testing problem

When we set a p-value threshold of, for example, 0.05, we are saying that there is a 5% chance that the result is a false positive. In other words, although we have found a statistically significant result, in reality, there is no difference in the group means. While 5% is acceptable for one test, if we do lots of tests on the data, then this 5% can result in a large number of false positives. For example, if there are 200 spots on a gel and we apply an ANOVA or t-test to each, then we would expect to get 10 false positives by chance alone. This is known as the multiple testing problem.

Multiple testing and the False Discovery Rate

While there are a number of approaches to overcoming the problems due to multiple testing, they all attempt to assign an adjusted p-value to each test, or similarly, reduce the p-value threshold. Many traditional techniques such as the Bonferroni correction are too conservative in the sense that while they reduce the number of false positives, they also reduce the number of true discoveries. The False Discovery Rate approach is a more recent development. This approach also determines adjusted p-values for each test.

However, it controls the number of false discoveries in those tests that result in a discovery (i.e. a significant result). Because of this, it is less conservative that the Bonferroni approach and has greater ability (i.e. power) to find truly significant results.

Another way to look at the difference is that a p-value of 0.05 implies that 5% of all tests will result in false positives. An FDR adjusted p-value (or q-value) of 0.05 implies that 5% of significant tests will result in false positives. The latter is clearly a far smaller quantity.

q-values

q-values are the name given to the adjusted p-values found using an optimized FDR approach. The FDR approach is optimized by using characteristics of the p-value distribution to produce a list of q-values. In what follows I will tie up some ideas and hopefully this will help clarify some of the ideas about p and q values. It is usual to test many hundreds or thousands of spot variables in a proteomics experiment. Each of these tests will produce a p-value. The p-values take on a value between 0 and 1 and we can create a histogram to get an idea of how the p-values are distributed between 0 and 1. Some typical p-value distributions are shown below. On the x-axis we have histogram bars representing p-values. Each has a width of 0.05 and so in the first bar (red or green) we have those p-values that are between 0 and 0.05. Similarly, the last bar

represents those p-values between 0.95 and 1.0, and so on. The height of each bar gives an indication of how many values are in the bar. This is called a density distribution because the area of all the bars always adds up to 1. Although the two distributions appear quite different, you will notice that they flatten off towards the right of the histogram. The red (or green) bar represents the significant values, if you set a p-value threshold of 0.05.

If there are no significant changes in the experiment, you will expect to see a distribution more like that on the left above while an experiment with significant changes will look more like that on the right. So, even if there are no significant changes in the experiment, we still expect, by chance, to get p-values < 0.05. These are false positives, and shown in red. Even in an experiment with significant changes (in green), we are still unsure if a p-value < 0.05 represents a true discovery or a false positive. Now, the q-value approach tries to find the height where the p-value distribution flattens out and incorporates this height value into the calculation of FDR adjusted p-values. We can see this in the histogram below. This approach helps to establish just how many of the significant values are actually false positives (the red portion of the green bar). Now, the q-values are simply a set of values that will lie between 0 and 1. Also, if you order the p-values used to calculate the q-values, then the q-values will also be ordered.

To interpret the q-values, you need to look at the ordered list of q-values. There are 839 spots in this experiment. If we take spot 52 as an example, we see that it has a p-value of 0.01 and a q-value of 0.0141. Recall that a p-value of 0.01 implies a 1% chance of false positives, and so with 839 spots, we expect between 8 or 9 false positives, on average, i.e. 839*0.01 = 8.39. In this experiment, there are 52 spots with a value of 0.01 or less, and so 8 or 9 of these will be false positives. On the other hand, the q-value is a little greater at 0.0141, which means we should expect 1.41% of all the spots with q-value less than this to be false positives. This is a much better situation. We know that 52 spots have a q-value less than 0.0141 and so we should expect 52*0.0141 = 0.7332 false positives, i.e. less than one false positive. Just to reiterate, false positives according to p-values take all 839 values into account when determining how many false positives we should expect to see while q-values take into account only those tests with q-values less the threshold we choose. Of course, it is not always the case that q-values will result in less false positives, but what we can say is that they give a far more accurate indication of the level of false positives for a given cutoff value. When doing lots of tests, as in a proteomics experiment, it is more intuitive to interpret p and q values by looking at the entire list of values in this way rather that looking at each one independently. In this way, a threshold of 0.05 has meaning across the entire experiment. When deciding on a cut-off or threshold value, you should do this from the point of view of how many false positives will this result in, rather than just randomly picking a p- or q-value of 0.05 and saying that everything with a value less than this is significant.

Microbiome analysis review

1_QIIME / MACQIIME (Clean sequences > Phylogeny tree > OTU Table)

2_Alpha / Beta diversity / PCA Differences in Microbiome community composition

3_Differential abundance on pre-specified taxa (SAS / GEE)

4_Ranked differential abundance on rarefied OTU table (R / edgeR)

5_Ranked differential abundance on non-rarefied OTU table (Phyloseq / DESeq2)

6_Ranked differential abundance on functional pathways (Picrust / LEFse & LDA)

Statistical analysis Summary Review (McMurdie & Holmes, 2014)

In recent generation DNA sequencing the total reads per sample (library size; sometimes referred to as depths of coverage) can vary by orders of magnitude within a single sequencing run. Comparison across samples with different library sizes requires more than a simple linear or logarithmic scaling adjustment because it also implies different levels of uncertainty, as measured by the sampling variance of the proportion estimate for each feature (a feature is a gene in the RNA-Seq context, and is a species or Operational Taxonomic Unit, OTU, in the context of microbiome sequencing). In this article we are primarily concerned with optimal methods for addressing differences in library sizes from microbiome sequencing data.

Variation in the read counts of features between technical replicates have been adequately modeled by Poisson random variables. However, we are usually interested in understanding the variation of features among biological replicates in order to make inferences that are relevant to the corresponding population; in which case a mixture model is necessary to account for the added uncertainty. Taking a hierarchical model approach with the Gamma-Poisson which gives the negative binomial (NB) distribution has provided a satisfactory fit to RNA-Seq data, as well as a valid regression framework that leverages the power of generalized linear models.

However, the variance for the negative binomial distribution becomes poisson when $\theta = 0$. Recognizing that $\theta > 0$ and estimating its value is necessary in gene-level tests in order to control the rate of false positive genes. Many false positive genes appear significantly differentially expressed between experimental conditions under the assumption of a Poisson distribution, but are nevertheless not-significant in tests that account for the larger variance that results from nonzero dispersion.

The uncertainty in estimating θ for every gene when there is a small number of samples — or a small number of biological replicates — can be mitigated by sharing information across the thousands of genes in an experiment, leveraging a systematic trend in the mean-dispersion relationship. This approach substantially increases the power to detect differences in proportions (differential expression) while still adequately controlling for false positives.

Although DNA sequencing-based microbiome investigations use the same sequencing machines and represent the processed sequence data in the same manner — a feature-by-sample contingency table where the features are OTUs instead of genes — to our knowledge the modeling and normalization methods currently used in RNA-Seq analysis have not been

transferred to microbiome research. Instead, microbiome analysis workflows often begin with an ad hoc library size normalization by random subsampling without replacement, or so-called rarefying.

Rarefying is most often defined by the following steps.

1. Select a minimum library size, NL,min. This has also been called the rarefaction level

2. Discard libraries (microbiome samples) that have fewer reads than NL,min.

3. Subsample the remaining libraries without replacement such that they all have size NL,min.

Often NL,min is chosen to be equal to the size of the smallest library that is not considered defective.

To our knowledge, rarefying was first recommended for microbiome counts in order to moderate the sensitivity of the UniFrac distance to library size, especially differences in the presence of rare OTUs. We demonstrate the applicability of a variance stabilization technique based on a mixture model of microbiome count data. This approach simultaneously addresses both problems of (1) DNA sequencing libraries of widely different sizes, and (2) OTU (feature) count proportions that vary more than expected under a Poisson model. We utilize the most popular implementations of this approach currently used in RNA-Seq analysis, namely edgeR and DESeq, adapted here for microbiome data. This approach allows valid comparison across OTUs while substantially improving both power and accuracy in the detection of differential abundance.

Suppose we want to compare two different samples, called A and B, comprised of 100 and 1000 DNA sequencing reads, respectively. In statistical terms, these library sizes are also equivalent to the number of trials in a sampling experiment. In practice, the library size associated with each biological sample is a random number generated by the technology, often varying from hundreds to millions.

Formally comparing the two proportions according to a standard test could technically be done either using a x2 test (equivalent to a two sample proportion test here) or a Fisher exact test. By first rarefying (Figure 1, Rarefied Abundance section) so that both samples have the same library size before doing the tests, we are no longer able to differentiate the samples (Figure 1, tests). This loss of power is completely attributable to reducing the size of B by a factor of 10, which also increases the width of the confidence intervals corresponding to each proportion such that they are no longer distinguishable from those in A even though they are distinguishable in the original data. This is because the variance of the proportion's estimate is multiplied by 10 when the total count is divided by 10. This is a common occurrence and one that is traditionally dealt with in statistics by applying variance-stabilizing transformations.

Statistical approach 1 (distinguishing patterns of relationships between whole microbiome samples)

The main goal is to distinguish patterns of relationships between whole microbiome samples through normalization followed by the calculation of sample-wise distances. Many early microbiome investigations are variants of this example, and also used rarefying prior to calculating UniFrac distances.

Microbiome studies often graphically represent the results of their pairwise sample distances using multidimensional scaling (also called Principal Coordinate Analysis, PCoA), which is useful if the desired effects are clearly evident among the first two or three ordination axes. In some cases, formal testing of sample covariates is also done using a permutation MANOVA (e.g. vegan::adonis in R) with the (squared) distances and covariates as response and linear predictors, respectively.

Normalizations Methods for distinguishing patterns of relationships between whole microbiome samples

1. DESeqVS. Variance Stabilization implemented in the DESeq package.

2. None. Counts not transformed. Differences in total library size could affect the values of some distance metrics.

3. Proportion. Counts are divided by total library size.

4. Rarefy. Rarefying is performed as defined in the introduction, using rarefy_even_depth implemented in the phyloseq package, with NL,min set to the 15th-percentile of library sizes within each simulated experiment.

5. UQ-logFC. The Upper-Quartile Log-Fold Change normalization implemented in the edgeR package, coupled with the topMSD distance (see below).

Distance Metrics Methods. for distinguishing patterns of relationships between whole microbiome samples. For each of the previous normalizations we calculated sample-wise distance/dissimilarity matrices using the following methods, if applicable.

1. Bray-Curtis. The Bray-Curtis dissimilarity first defined in 1957 for forest ecology.

2. Euclidean. The euclidean distance treating each OTU as a dimension.

3. PoissonDist. Our abbreviation of PoissonDistance,a sample-wise distance implemented in the PoiClaClu package.

4. top-MSD. The mean squared difference of top OTUs, as implemented in edgeR.

5. UniFrac-u. The Unweighted UniFrac distance.

6. UniFrac-w. The Weighted UniFrac distance.

Statistical Approach 2 (differential abundance analyses)

The goal is to detect microbes that are differentially abundant between two pre-determined classes of samples. This experimental design appears in many clinical settings (health/ disease, target/control, etc.), and other settings for which there is sufficient a priori knowledge about the microbiological conditions, and we want to enumerate the OTUs that are different between these microbiomes, along with a measure of confidence that the proportions differ.

Normalization methods for differential abundance analyses.

For each simulated experiment, we used the following normalization/ modeling methods prior to testing for differential abundance.

1. Model/None. A parametric model was applied to the data, or, in the case of the t-test, no normalization was applied (note: the t-test without normalization can only work with a high degree of balance between classes, and is provided here for comparison but is not recommended in general).

2. Rarefied. Rarefying is performed as defined in the introduction, using rarefy_even_depth implemented in the phyloseq package, with NL,min set to the 15th-percentile of library sizes within each simulated experiment.

3. Proportion. Counts are divided by total library size.

Testing methods for differential analyses. For each OTU of each simulated experiment we used the following to test for differential abundance.

1. two sided Welch t-test. A two-sided t-test with unequal variances, using the mt wrapper in phyloseq of the mt.maxT method in the multtest package.
2. edgeR - exactTest. An exact binomial test (see base R's stats::binom.test) generalized for overdispersed counts and implemented in the exactTest method of the edgeR package.
3. DESeq - nbinomTest. A Negative Binomial conditioned test similar to the edgeR test above, implemented in the nbinomTest method of the DESeq package.
4. DESeq2 - nbinomWaldTest. A Negative Binomial Wald Test using standard maximum likelihood estimates for GLM coefficients assuming a zero-mean normal prior distribution, implemented in the nbinom WaldTest method of the DESeq2 package.

All tests were corrected for multiple inferences using the Benjamini-Hochberg method to control the False Discovery Rate. Please note that in the context of these simulations library size is altogether different from effect size; the former being equivalent to both the column sums and the number of reads per sample.

Biostatistics Inference Study Guide

Goals of the class

Little attention is paid to understanding the underlying statistical models we use

Parameter estimated, p-values, CI's seem to just pop out of the statistical software

Need to understand the implication of an assumed probability model to assess if a model is reasonable

Review the basic concepts of probability and statistical models

Discuss how we can in general obtain estimates of parameters, construct hypothesis tests, and obtain confidence intervals

LECTURE 1

Statistical procedures to compare continuous outcome between two groups: T-test / Mann-Withney-U for non parametric data

Statistical procedures to compare continuous outcomes between multiple groups: ANOVA / Kruskall-Wallis for non parametric data

#Statistical procedures to compare categorical outcomes between different categorical groups: Chi Square / Fischer's Exact Test

Statistical procedures to relate a continuous covariate to a continuous outcome: Pearson correlation / Spearman correlation (non parametric rank correlation) / Linear regression

Statistical procedures to relate a continuous outcome after adjusting for other covariates: GLM (multiple regression)

Statistical procedures to relate a covariate to binary outcomes after adjusting for other covariates: Logistic Regression /Probit Regression

Key Assumptions

To ESTIMATE Bo and B1 and make inferences on those parameters we must make some assumptions. These typically include the following which I like to summarize with the acronym LINE.

- Linearity (Assume a linear relationship between Y and X)

- Independence: The observations Y are independent from one another

- Normality: the residuals about the trendline are normally distributed

- Equal Variance: the variance of the response at each X does not change

Standard statistical procedures require making certain assumptions

Those assumptions are really about a statistical model. In some cases they are not reasonable for our data and we need a new model. Learn how we go from statistical modeling and data to estimates, tests and p-value (how the computer gets these numbers in our output).

Lecture 2

Experiment and sample space

- Definition of Experiment: An experiment is any action or process whose outcome is subject to uncertainty

- Definition of Sample Space: The sample space of an experiment, denoted by {S}, is the set (or collection or enumeration or list) of all possible outcomes of that experiment

- The list must be mutually exclusive and exhaustive

Events

- Definition of Event: An event is any collection (subset) of outcomes contained in the sample space (S)

1) Simple: consists of exactly one outcome

2) Compound: consists of more than one outcome

Three axioms of probability:

- For any event A such as X = 3, P(A)>= 0
- P (S) =1
- If A1, A2, A3 ... Ai are an infinite collection of disjoint events (no common outcome) then P (A1 or A2 or ? or Ai) is SUM (I 1-> OO) P(Ai)

#Probability rules summary

- # P (A or B) = P(A) + P(B) - P(A and B)
- # P (not A) = 1 - P(A)
- # P (A|B) = P(A and B) / P(B): This is the multiplication rule
- # Extension of the Multiplication rule: P(A|B)*P(B) = P(A and B) = P(B|A)*P(A)
- # Therefore, we can take it one step further and flipping the Conditional: P(A|B) = [P(B|A)*P(A)]/P(B)
- # P (A and B and C) = P(A|B and C)*P(B|C)*P(C): Note that the order in which A, B and C are used is exchangeable!
- # Law of total probability: for all Ai which are disjoint events, P(B) = SUM [P(B|Ai)*P(Ai)]
- # Bayes Rule: By combining the flipping the conditional rule and the law of total probability, P[Ai|B] = [P(B|Ai)*P(Ai)] / SUM [P(B|Ai)*P(Ai)]
- # If A and B are independent P(A and B) = P(A)*P(B) and P(A|B) = P(A)
- # If A and B are disjoint, P(A and B) = 0 Since they cannot exist together
- # Therefore If A and B are disjoint they cannot be independent, because P(A and B)= P(A)*P(B) =/= 0

 # Life Tables: typically deal with three probabilities

1) The probability an individual dies in the t^th interval given that they have survived to the beginning of that interval P(T = t |T > t - 1) = # dead during interval / # Alive at the begining of the interval

2) The probability a person survives the t^{th} interval given that they have

survived to the beginning of that interval P(T > t|T > t-1) = 1 - [= # dead during interval / # Alive at the begining of the interval]

3) The unconditional probability of surviving past the t^{th} interval P(T > t) = Product of all previous intervals in (2) because it is the multiplication rule for independent Events" P(A and B and C) = P(A)*P(B)*P(C)

Inconsistent Follow-up

- A person is considered "at risk" for failure in an interval t if they have not yet experienced an event and are not censored by the start of the interval.

- If we know the number of people who start each interval and the number who die in each interval, then we can easily calculate P(T = t|T > t -1 and C>t-1) as the proportion of failures among those at risk.

- Assume that P(T = t|T > t -1 and C>t-1)= P(T = t|T > t -1). That is we assume conditional independence!

- The conditional survival is straightforward as well: P(T > t | T > t - 1) = 1 - P(T = t|T > t - 1) because P(T=/=t|T>t-1) = 1-P(T=t|T>t-1) = P(T>t|T>t-1)

Definition of Random Variable

For a given sample space S of some experiment, a random variable (rv) is any rule that associates a number with each outcome in S . In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers.

****Discrete Random Variable****

- A discrete random variable is an rv whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on.

****Continuous Random Variable****

- A random variable is continuous if both of the following apply:

1) Its set of possible values consists either of all numbers in a single interval on the number line (possibly infinite in extent, e.g., from -00 to +OO) or all numbers in a disjoint union of such intervals.

2) No possible value of the variable has positive probability, that is, P(X = c) = 0 for any possible value c.

****Definition of Probability Mass Function****

The probability distribution or probability mass function (pmf) of a discrete rv is defined for every number x by p(x)=P(X=x). This is a step wise function

****Definition of Cumulative Distribution Function****

The cumulative distribution function (cdf) F(x) of a discrete rv X with pmf p(x) is defined for every number x by F(x) = P(X =< x)

PMF and CDF Properties

- Recall that (X=x) is just an event; that is [s in S: X(s)=x].

- In particular what is the range of possible values for p(x) for all x? 0<p(x)<1

- What must sum_x p(x) equal? Sum_x p(x) = 1

PMF and CDF Properties

- Similarly, X =< x is also an event

- F(-OO) = 0 and F(+OO) = 1

- Note that X =< x is the complement of X > x. P(X =< x) = 1 - P(X>x)

- If x < y, F(x) < F(y): We say that F(x) is Right continuous

****Definition of Expectation****

Let X be a discrete rv with set of possible values D and pmf p(x).

The expected value of any function h(X), denoted by E{h(X)} sum{x in D} h(x)*p(x)

Note: This expected value will exist provided that sum_{h(x) in D} |h(x)|p(x) < +OO

Expectation Properties

- E(b)=b if b is a constant

- E(aX)=aE(X) where a is a constant

- E(sum of X) = sum of E (X)

- If h(X) =aX+b then E{h(X)}= aE(X)+b

- In general, h{E}(X) ne {E}{h(X)}. However, if h(X) is a convex function then h{E(X)} =< E{h(X)} (Jensen's Inequality)

****Definition of Variance****

Let X have pmf p(x) and expected value mu. The variance of X denoted by V(X) or sigma^2 is

V(X)=E{(X-mu)^2}=sum(x-mu)^2*p(x)

- The standard deviation sigma is just sqrt{V(X)}.

- The variance is just the expectation of a function which we know how to compute

Common PMF Families

Bernoulli Experiment and Random Variable

A fancy name for a very simple experiment. A Bernoulli experiment meets the following two conditions

1) The experiment consists of a single trial which can result in only one of two outcomes, which we generically call a success (S) and failure (F)

2) The probability of success is p

A Bernoulli random variable X associated with this experiment is the indicator for whether or not the experiment was a success; that is

X(S)=1 and X(F)=0.

Bernoulli Random Variable

The pmf of a Bernoulli random variable is p(x;p) = p if x=1, 1-p if x=0, and 0 otherwise

The cdf of a Bernoulli random variable is F(x;p) = 0 if x<0, 1-p if 0 =< x < 1, and 1 if X >= 1

Bernoulli Random Variable

- E(X) = p

- V(X)= p(1-p)

Binomial Experiment and Random Variable

A binomial experiment meets the following four conditions

1) The experiment consists of n trials ***where n is fixed in advance***

2) Each trial can result in only one of two outcomes, which we generically call a success (S) and failure (F)

3) The trials are independent

4) The probability of success is constant from trial to trial and equal to p

5) A binomial random variable X associated with this experiment is the total number of success.

6) Connection between Bernoulli and Binomial Random Variables: A binomial random variable is the sum of n Bernoulli random variables

Binomial Expectation and Variance

- E(X) = np

- V(X) = np(1-p)

Poisson Random Variable

- Poisson distribution with parameter Y > 0

- E(X) = Y, V(X) = Y

- Cdf is P(X<A) = ppoison (A,Y), pmf is P(X=A) = dpoisson (A,Y)

Continuous Random Variables

Discrete Random Variables: Binomial, Bernoulli, Poisson (pmf and cdf are non continuous)

Continuous Random Variables: The set of possible values includes all numbers in a single interval on the number lines, and P(X=c) = 0): Therefore we have to look at the density of the observations)

Review of integration

- Integral of f(x)dx from a to b is the sum of the area under the curve from a to b

- Area = length x height. here length is the bin width and the height is determined by the probability function f(x): Therefore density = probabilty/bin width

Probability density function

- Probability density function for any two numbers a < b is P(a<X<b) = Integral of f(x)dx from a to b: This means that the probability of X being in between a and b is the area under the curve from a to b. Remember that P(X=c) = 0 for a single point!

Cumulative distribution function

- F(a) = P(X<a) = integral from -OO to a of f(y)dy

- to go grom the cdf to the pmf, use P(a<X<b) = F(b) - F(a)

- pdf to cdf = integration cdf to pdf = derivation

definition of percentile

- for 0<p<1, let np = the 100p^th percentile defined by integral from -OO to np of f(y)dy = p

- the 50th percentile is 100p^th = 50 so p = 0.5 and solve integral of f(y)dy = 0.5

Solving continuous random variables in R

- d[distribution name] (x, parameter) = pmf/pdf

- p[distribution name] (x, parameter) = cdf

- q[distribution name] (x, parameter) = 100p^th percentile

Expectation and Variance of a random Variable

- E(x) = integral -OO to OO of x*f(x)dx

- V(x) = E(x^2) - [E(x)]^2

Normal distribution

- Known as the gaussian distribution, bell shaped, symmetric so mean = median, support (-OO to OO)

- Distribution is determined by mean u and standard deviation sd

- the percentiie at z for all normal distributions can be expressed as u+z*sd

Normal distribution example

- X ~ normal (u=8.8, sd = 2.8)

- P(X>10) = 1 - P(X<10) = 1 - pnorm(10, 8.8, 2.8)

- P(5<X<10)= pnorm (10, 8.8, 2.8) - pnorm (5, 8.8, 2.8)

- What value has 5% of the distrubution underneath it P(X<a) = 0.05 then qnorm (0.05, 8.8, 2.8)

Empirical Rule

- 1SD = 68%, 2SD = 95%, 3SD = 99.7%

Normal Distribution Approximation

- X is approximately normal for a binomial distribution with np>10 and n(1-p)>10

- X is approximately normal for a poisson distribution with mean Y and sd sqrt(Y)

Uniform distribution

- Used when each number of the finite interval has an equal chance of occuring, (ie, waiting time for the bus)

- X is siad to have a uniform distribution over [A,B]

- f(x) = 1/(A-B) for A<X<B

- Mean = (A+B)/2, Variance =)1/12(b-a)^2, 100p^th percentile = p(B-A) + A

Gamma function

- Beta is known as the scale parameter (>0), and alpha is known as the shape parameter (>0)

- E(X) = alpha*Beta, V(X) = alpha*Beta^2

Exponential distribution

- if X follows a gamma distribution with alpha = 1 and Beta = 1/Y then the exponential distribution

- f(x) = Y*exp(-Y*x) for the pdf

- F(x) = 1-exp(Y*x) for the cdf

- E(X) = 1/Y, V(X) = 1/Y^2

Chi square distribution

- Let X follow a gamma distribution with alpha = mu/2 and beta = 2: X is said to follow a chisquared distribution with degrees of freedom mu

- E(X) = mu, V(X) = 2*mu

- If Z follows a standrand normal distribution, then Z^2 follows a chisquare distribution with 1 degree of freedom (mu = 1)

Beta distribution

-if alpha = Beta then the distribution is symmetric

- if alpha = beta = 1 then the distribution is uniform

- E(X) = alpha / (alpha + beta), V(X) = alpha * beta / (alpha + beta)^2 * (alpha + beta + 1)

Multivariate Distributions

- Joint Probability Mass Function: Let X and Y be two discrete random variables defined on the sample space S. p(x,y) is defined for each pair of numbers (x,y) is p(x,y) = P(X=x, Y=y. p(x,y) > 0 for all x and y. sum over x AND sum over y of p(x,y) = 1

- Joint probability density function: Let X and Y be continuous random variables. f(x,y) is the joint probability function for X and Y if for any dimensional set A P(X,Y) = integral over A f(x,y)dydx

Marginal Probability mass function

- the marginal probability mass function of X and Y, denoted by Px(x) and Py(Y) is denoted by summing over the OTHER variable on which we do not want the marginal probability. i.e. px(x) = Sum over y p(x,y) and py(Y) = sum over x p(x,y)

independence example for a joint probability function

- X~exp (Y): Fx(xi) = Yexp(-Yxi) (cdf formula); the joind distribution for all xi is the product of Fx(xi) = Yexp(-Yx1)*Yexp(-Yx2)*...*Yexp(-Yxn). Given that expa*expb = exp (a+b), Fx(xi) becomes Y^n*exp(-Y*sumYi)

- P(Z>t) = P(X1>t and X2>t and ... and Xn>t): since they are independent P(X>t) = P(X>t)^n. P(X<t) = cdf = 1-exp(-Yt) so P(Z>t) = [1-cdf]^n = [exp(-Yt)]^n

Expected Value

- let X be height and y be weight. BMI = H(X,Y) = Y/X^2 = (Y) * (1/X^2) = h(x)*h(y) so p(x,y) = px(x)*py(y). E[H(x,y)] = sum over X AND sum over Y h(x,y)*p(x,y): E[h(x,y)]= sum over X h(x)*p(x) * sum over Y h(y)*p(y) = E(h(x))*E(h(y))

Covariance

- Cov (X,Y) = E(XY)- E(X)E(Y)

- Cov (X+Y, Z) = Cov (X,Z) + Cov (Y,Z)

- Cov (aX, bY) = ab*cov(X,Y)

- Cov (X+c, Y) = Cov (X, Y)

Correlation

- Corr (X,Y) = Cov (X,Y)/sd(x)*sd(y)

- -1 < corr (X,Y) < 1

- If X and Y are independent then corr (X,Y) = 0, but the reverse is not necessarily true!

- Correlation is a measure of linear association

Conditional Distributions

- P(A|B) = P(A and B) / P(B) so conditional probability = joint probability / marginal probability : p y|x (y|x) = p(x,y) / px(x) and f y|x (y|x) = f(x,y) / fx(x)

- Conditional mean uY|X = x = E(Y|X=x) (Sum of y*py|x(y|x) if X and Y are discrete, Integral from -OO to OO if X and Y are continuous)

- Conditional Variance sigma Y|X=x = E(Y^2|X=x) - u^2 Y|X=x

Example

- X ~ Uniform (0,1)

- Y|X ~ Uniform (0,X)

- f(x) = 1/(B-A) if 0<X<1, O otherwise. f(y|x) = 1/(X-0) if 0<Y<X, 0 otherwise.

- since conditional probability = joint probability / marginal probability and we know f(y|x) and f(x), f(x,y) = f(y|x)*f(x). f(x,y) = (1/X)*(1/(1-0)) if 0<Y<X<1

- fy(y) = integral over all X of f(x,y) = integral of (1/x)dx from y to 1 since 0<Y<X<1 = ln(1)-ln(y) = -ln(y) for 0<y<x

Bivariate normal density

- conditional x|y is normal with mean ux|y = ux + ro*sx*(y-uy/sy) and variance sx|y = s^2(x)(1-ro^2)

- for a conditional mean x, ux|y > ux if y>uy, because (y-uy/sy) > 0

- for a conditional variance x, sx|y < sx because since ro<1, 1-ro^2 less than 1