Frequentist approach (p value)

1. Stating a hypothesis
2. Collecting data
3. Computing a summary statistic
4. Test hypothesis (no)->test/infer how frequently we would observe data by chance akine if the null hypothesis is true
    a. P value how likely you are to observe that difference IF the null hypothesis is true

- Precision: repeated estimates give similar results (dots end up in similar place on target)
- Accuracy: how close they are to the true value

Probability

- (P) of an event is the number of times the event will occur (a) divided by the total number of possible events (n)
    - P=a/n
- Product rule: multiply 2 individual probabilities tg is equal to the product of their individual probabilities
- Sum rule: the probability of 2 or more mutually exclusive events occurring is equal to the sum of their individual probabilities
- conditional probability: probability of an event given that another event has happened
    - Notated by |
    - Used whenever considering events that are not independent

Statistical measures and distributions

- Tests of difference
    - Parametric
    - Non parametric
- Means
    - Arithmetic mean
    - Harmonic mean
    - Geometric mean
- Variation
    - Sample variance
    - Stdv (sqrt of variance)
    - Standard error (standard deviation divided by n)

Probability distribution

- Binomial: when plotted out=binomial density function
    - N: number of trials
    - X: number of successes
    - P
    - Q

Max likelihood approach

1. Collecting data
2. Developing model with parameters
3. Inputting raw data into model
4. Computing likelihood of data for all possible parameter values

Bayesian approach

- Incorporates prior info to compute a probability estimate
- Directly yields probability of hypothesis being true (posterior probability)

4.3 Problems with P values

- More difficult to interpret via frequentist approach
    - Do not measure the probability that the hypothesis is true
    - Interpreted as the chance that if the null hypothesis is true, you will get a similar or more extreme result if you repeat the experiment many times.
- Overstates strength of evidence
    - If P value is low but not significant: researchers may not reject null hypothesis
    - File drawer effect: negative results often end up unpublished

Solutions to listed problems

- Bayes factor: can replace P value to quantify strength of hypothesis
- Supplement P values with estimates of confidence in the P value and assume that it is not a false positive
- Supplement p value by reporting effect sizes and confidence

5-Maximum likelihood

- Max likelihood (ML): estimates parameter value that maximizes the probability of obtaining the observed data under a given model-use max amount of info from data
- Advantages
    - Model based=easy comparison
    - Use data in raw form (not summary statistics)
    - Accurate and precise
- Disadvantages: require large sample sizes
    - Smaller sample sizes will produce biased and less precise results

6. Bayesian approaches and markov chain monte carlo

- Bayesian inference differences:
    - Probabilities defined and interpreted differently: yield more direct probability answer easier to interpret than a P value
        - P=0.95=95% probability that hypothesis is true
    - Can include prior data when estimating the posterior probability that a hypothesis is correct (ex. Multiplying the likelihood function by the prior information) examples include:
        - Can be used in estimating effective population size when population census size is known

        - Models that incorporate mutation dynamics
    - Advantage: facilitates decision making when we want to integrate all available knowledge
    - Disadvantage: can be strongly influenced by prior data=less objective
        - 2 ppl can use diff prior info=each get different results HOWEVER we can quantify effects of different priors

6.1 Markov chain Monte Carlo (MCMC) EXAMPLE: FIG A7

- Method for simulating random samples from a probability distribution
    - Way to get an estimate of posterior mean and interval parameters
        - Used to sample from posterior distribution of a parameter to generate probability estimate of said parameter
    - Uses **markov chain**: generates series of random variables whose future state depends only on the current state at any point in the chain
- Combines:
    - Markov chain model (chain of random steps
    - **Monte carlo proces**s: draw a random number necessary at each step
        - **Burn in:** first few thousands of steps, later discarded to reduce influence of the starting point
            - Once burnt in=**converge**d simulation: independent of starting point
    - Disadvantages
        - We don't know if we did enough burn in steps (to avoid bias)
        - Difficult to code=errors likely/hard to detect

7. Approximate bayesian computation (ABC)

- Uses prior data to output approx posterior probability distribution
    - Posterior approximated by summarizing the data using multiple summary statistics
    - Steps: (also called summary statistic matching)
        1. Replace (summarize) raw observed data with multiple summary statistics
        2. Compute the same summary statistics for population models under consideration
        3. Match observed summary stats to those from simulated populations to choose population parameter that is best fit
- Advantages
    - Use nearly all info from data
    - Not as computationally demanding than bayesian methods
        - Take hours-days vs weeks
    - Can be used in large datasets
    - Allow comparisons of most demographic scenarios that can be simulated
    - Estimate key parameters of model (ex.bottleneck minimum effective size)
    - Define priors

- Disadvantages
    - Question accuracy
    - Compatibility issues

8. Parameter estimation, accuracy, and precision-Which estimator and approach perform best?

- Performance depends on question, sample size, sample characteristics of parameter being estimated, and effect size.
    - Efficiency: refers to ability to extract info from the data and achieve high accuracy and precision
        - Accuracy: tendency to yield estimates near the true population parameter value
            - Genomics can improve accuracy
                - SNPs: improve inbreeding estimation=detect inbreeding depression (precise estimation)
                - Loci on chromosomes
- Use different estimators when assessing a given question:
    - Mean and median
    - Moment basedad likelihood based estimators
- Random and representative sampling: important so estimate is biased

9. Performance evaluation

- Quantification of the accuracy, precision, power, and robustness of a statistical estimation
- Use stat methods without risking making bad management decisions BUT its never conducted thoroughly
- 4 steps:
    1. Generate test dataset with known parameter value for a parameter of interest
    2. Estimate parameter
    3. Repeat steps 1 and 2 1,000 times
    4. Compute proportion of 1,000 estimates that give the true parameter most accurately and precisely

10. Coalescent and genealogical information (Box A2 gives example)

- **Box A3 shows it being used in frequentist, likelihood and bayesian approaches**
- Colesce: fuse, unite, come together
    - Process of tracing backward thru time and join haplotypes from different individuals in same parent or ancestor
    - Why use it?
        - Estimate population genetic parameters and infer demographic status
    - Yields a distribution of times to the most recent common ancestor between gene copies in a genealogy
    - Can be used w frequentist, ML, bayesian approaches to:
        - Generate expected distribution of allele frequencies to test parameters

    - Allows us to see genealogical info, from DNA sequences, at a locus
- Singletons: rare allele found as a single copy
- Must sample many genes to get accurate+precise estimates of populations demographic history

Chapter 5: random mating populations/Hardy-Weinberg principle

- Models
    - essential to understanding genetics of natural pops. Models:
        - Make us define parameters that need to be considered
        - Allow us to test hypotheses
        - Allow us to generalize results
        - Allow us to predict how a system will operate in the future
    - Should be as simple as possible bc they need to be able to be tested and rejected (easier with a simple model), and simple models are more general and therefore more applicable to wider number of situations
    - Hardy-Weinberg (HW) equilibrium: allele and genotype frequencies will remain the same from generation to generation
        - Based on Mandelian segregation for diploid organisms that are reproducing sexually in combination with principles of probability
    - Assumptions when making our model:
        - 1. Random mating
        - 2. No mutation
        - 3. Infinite population size (no genetic drift, no loss of genetic material over time)
        - 4. No natural selection
        - 5. No immigration/gene flow
        - Consequences of making these assumptions:
            - This population will not evolve (all alleles have equal probability of inheritance)
            - Genotype frequencies will be in binomial (HW) proportions
        - HW allows us to describe a pop by frequencies of alleles at a locus rather than the many diff genotypes that can occur at a single locus
    - HW proportions
        - Heterozygote = Aa (N12); homozygote = AA(N11) or aa (N22); total individuals = N
        - Allele frequencies are:
            - p = freq(A) = (2N
        - HW proportions are not applicable to real life– there's too much variation and we make too many assumptions in this model to account for all of them

        - Chi-square test can be used to determine if the observed difference in expected HW genotypes is greater than what we would expect by chance alone
    - Small sample sizes introduce systematic bias when trying to predict genotypes– the less individuals, the less chances of different genotypes, but this doesn’t reflect the actual population
    - HW principle can be used to to estimate allele frequencies at loci in which there is not a unique relationship between phenotype and genotype

Chapter 5

Mendels laws of inheritance

1. Dominance and uniformity
2. Segregation
3. Independent assortment

Basics

- Allele
- Genotype

Models

- Conceptual
- Mathematical
- Pros
- Cons

Hardy weinberg

- Conditions
    - Random mating
    - No mutation
    - Infinite population size (no genetic drift)
    - No natural selection
    - No immigration/gene flow
    - All of these factors result in
- Used in practical applications
    - Determine probability
    - Compare observed and expected
        - If they're not what we would expect:
            - Alternate modes of inheritance at play (sex linkage)
- Limitations
    - Small sample size: don't use chi-squared test when any expected number is less than 5

Alleles and distinguishing genotype from phenotype

Sex linkage

- Responsible for the greater occurrence of recessive disorders and recessive traits observed in heterogametic sex (males, XY). many sex determining mechanisms that exist
- Pseudoautosomal regions (PAR): regions on sex determining chromosomes that carry functional genes present on both variations of the sex chromosome
    - Relatively small in humans and fruit flies compared to other species

    - Sometimes genotypes of either sex need to be examined individually to help identify PAR

Genetic variation ->equation

- Average expected heterozygosity
    - Calculated by subtracting homozygosity from 1.0 at n loci within a population provides the best estimate of genetic variation
- Total # of alleles at a locus is a good measure of genetic variation and can be used to complement H
-

Discussion questions:

1. Is statistically power actually a good power analysis
    - Statistical power: ability to reject the null hypothesis, based on effect size, a value, and sample size
    - Power analysis: analyzing power of the stats
        - Ex. “i think the effect size will be this, how many samples do i need to maximize that effect”
        - Used when researchers dont get stat significant result
        - Should be done before

2. HW proportions and equations-isnt it the same
    - Expectations of HW vs equilibrium state
        - Were not testing for that condition bc its impossible

3. Does this mean that the question “which multiple testing correction methods should be used” is redundant? Or does it matter what data you have or conclusions/hypothesis you have (what is the point in correction)
    a. Doesn't matter how you try to correct p values->didnt find the relationship they were looking for
    b. You can use either data correction method regardless

4. What are some ways basenyan model can be applied to research