Statistical Inference - Notes

Tanner Prestegard

Course taken from 6/1/2015 - 6/28/2015

Introduction

- Statistical inference: generating conclusions about a population from a noisy sample.

Probability

- Given a random experiement, a probability meausre is a population quantity that summarizes the randomness.
- Specifically, probability takes a possible outcome from the experiment and:
  - Assigns it a number betwee 0 and 1.
  - The probability that something happens should be 1.
  - The probability of the union of any two sets of outcomes that are mutually exclusive is the sum of their respective probabilities.
    - $P(A \cup B) = P(A) + P(B)$
- Rules that probability must follow:
  - The probability that nothing occurs is 0.
  - The probability that something occurs is 1.
  - The probability of something is 1 minus the probability that the opposite occurs.
  - The probability of at least one of two or more things that cannot simultaneously occur is the sum of their respective probabilities.
  - If an event A implies the occurrence of event B, then the probability of A occurring is less than the probability of event B.
  - For any two events, the probability that at least one occurs is the sum of their probabilities minus the sum of their intersection.
    - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Probability densities and mass functions for random variables are useful for modeling and thinking about probabilities for the numeric outcome of experiments.
- A random variable is the numerical outcome of an experiment.
  - Can be discrete or continuous.
- Probability mass function: a probability mass function evaluated at a value corresponds to the probability that a random variable takes that value. To be a valid PMF, a function must satisfy:
  - It must always be larger than or equal to zero.

  - The sum of the probabilities for all possible values of the random variable has to add up to one

- Example: Bernoulli distribution for a coin flip
  - X = 0 represents tails and X = 1 represents heads. Probability of getting heads is θ.
  - $p(x) = \theta^x (1-\theta)^{1-x}$

- A probability density function (PDF) is a function associated with a continuous random variable.
  - Must be larger than or equal to zero.
  - The total area under it must be one.
  - Areas under PDFs correspond to probabilities for that random variable.

- CDF and survival function.
  - The cumulative distribution function of a random variable X returns the probability that the random variable is less than or equal to the value x.
  - $F(x) = P(X \leq x)$
  - The survival function is just $1 - F(x)$ and gives the probability that X is larger than or equal to x.

- Quantiles
  - Sample quantiles - 95th percentile means that 95% people did worse and 5% did better.
  - Population quantiles: The αth quantile of a distribution with CDF $F(x)$ is the point x_α such that $F(x_\alpha) = \alpha$.
  - A percentile is simple a quantile with the quantile expressed as a percent.
  - The median is the 50th percentile.
  - R can approximate quantiles for you for common distributions.
    * Use the "q" commands: `qbeta, qnorm, qpois`, etc.

Conditional probability

- Let B be an event such that $P(B) > 0$.
- Then the conditional probability of an event A occurring is $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- If A and B are independent, $P(A|B) = P(A)$.
- Bayes' theorem: $P(A|B)\,P(B) = P(B|A)\,P(A)$
- Definitions: consider an example where B means you have a disease and A is a positive test result.
  - Sensitivity: $P(A|B)$
  - Specificity: $P(\text{not A}|\text{not B})$
  - Positive predictive value: $P(B|A)$
  - Negative predictive value: $P(\text{not B}|\text{not A})$
- Using Bayes' theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Independence: event A is independent of event B if $P(A \cap B) = P(A)\,P(B)$
- IID random variables: independent and identically distributed.
  - Independent: statistically unrelated from one to another.
  - Identically distributed: all having been drawn from the same population model.

Expected values

- The mean is a characterization of the center of a distribution.
  - $E[X] = \sum_x xp(x)$
- The variance and standard deviation are characteristics of how spread out a distribution is.
- For a continuous random variable, the expected value is again exactly the center of mass of the density.
- The average of random variables is itself a random variable and its associated distribution has an expected value, but the center of this distribution is the same as that of the original distribution.
  - The sample mean is **unbiased** because its distribution is centered at what it's trying to estimate.
  - To put it another way: the distribution of averages of samples will have the same mean as that of the random variable sample itself.
  - The more data that goes into the sample mean, the more concentrated its density will be around the population mean.

Introduction to variability

- Variance is a measure of the spread of a distribution.
  - $Var(X) = E\left[(X-\mu)^2\right] = E\left[X^2\right] - E[X]^2$
  - The square root of the variance is the standard deviation.
- Sample variance: average squared distance of the observations from the sample mean
  - $S^2 = \frac{\sum_{i-1}(X_i - \bar{X})^2}{n-1}$
  - The expected value of the sample variance is the population variance.
  - Dividing by $n-1$ instead of n is what makes this an unbiased estimate of the population variance.
- Standard error on the mean
  - Recall that the average of random samples from a population is itself a random variable.
  - Expected value of sample mean: $E\left[\bar{X}\right] = \mu$
  - Variance of sample mean: $Var\left(\bar{X}\right) = Var\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n^2}Var\left(\sum_i X_i\right) = \frac{1}{n^2}\sum_i \sigma^2 = \sigma^2/n$
- The standard deviation talks about how variable the population is.
- The standard error $S/\sqrt{n}$ talks about how variable averages of random samples of size n from the population are.
- The variance of a sample mean is σ^2/n, and we estimate it with S^2/n.
  - The standard error of the sample mean is $s/\sqrt{n}$.
- Chebyshev's inequality: the probability that a random variable X is at least k standard deviations from its mean is less than $1/k^2$
  - $Pr(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$

Some common distributions

- Binomial distribution
  - Comes from the Bernoulli distribution - result of a binary outcome like a coin flip. $P(X = x) = p^x (1-p)^{1-x}$
  - A binomial variable is the sum of many IID Bernoulli trials.
    - With n trials, $P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{1-x}$
- Normal distribution
  - With expected value μ and variance σ^2, the distribution is given by: $\left(2\pi\sigma^2\right)^{1/2} e^{-(x-\mu)^2/2\sigma^2}$
  - Standard normal distribution has $\mu = 0$ and $\sigma^2 = 1$.
  - We can convert to a standard normal distribution from another distribution by taking $X \to \frac{X-\mu}{\sigma}$.
  - We can go the other way by taking $X \to \mu + \sigma X$.
- Poisson distribution
  - Used to model counts, event time data, survival data, contingency tables, and more.
  - $P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$
  - Mean = variance = λ
  - Poisson approximation to the binomial:
    - $X \sim Binomial(n, p)$
    - $\lambda = np$
    - n gets large, p gets small.

Asymptotics

- Term for the behavior of statistics as the sample size or some other relevant quantity goes to infinity or zero.
- Very useful for simple statistical inference and approximations.
- Also form the basis for frequency interpretation of probabilities (the long run proportion of times an event occurs).
- The Law of Large Numbers (LLN)
  - The sample mean converges to the population mean in the limit of infinite trials.
- An estimator is **consistent** if it converges to what you want to estimate.
- The LLN says that the sample mean of iid samples is consistent for the population mean.
- Typically, good estimators are consistent - we should expect to get the right answer if we collect infinite data.
- The Central Limit Theorem (CLT)
  - The distribution of averages of iid variables (properly normalized) becomes that of a standard normal distribution as the sample size increases.
  - $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\text{Estimate - Mean of estimate}}{\text{Std. Err. of estimate}}$
  - The useful way to think about the CLT is that $\bar{X}_n$ is approximately $N\left(\mu, \sigma^2/n\right)$.
  - Standard deviation is equal to the standard error on the mean.

  - No guarantee that n is "big enough."

- Confidence intervals
  - For a sample mean of a normally distributed random variable, $\bar{X} \pm 2\sigma/\sqrt{n}$ is called a 95% confidence interval for μ.
  - Sometimes, if your number of trials isn't large enough for the CLT to be applicable, you can form the interval with $(X+2)/(n+4)$.
  - Taking the mean and adding and subtracting the relevant normal quantile times the standard error yields a confidence interval for the mean.
  - Confidence intervals get wider as the coverage increases.

T confidence intervals

- Using T quantiles rather than Z quantiles.
- Tails will be a little wider than for Z intervals.
- T intervals are useful for small sample sizes, will become like the Z intervals in the limit of lots of data.
- The T distribution is indexed by the degrees of freedom; it gets more like a standard normal as df gets larger.
- The T distribution assumes that the underlying data are IID Gaussian with the result that $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ follows Gosset's T distribution with $n-1$ degrees of freedom.
- The T intervals are $\bar{X} \pm t_{n-1} S/\sqrt{n}$, where t_{n-1} is the relevant quantile.
- This distribution works well whenever the distribution of the data is roughly symmetric and mound-shaped.
- Paired observations are often analyzed using the T interval by taking differences.
  - Get mean of differences: $\bar{Z}$
  - Get sigma of differences: $s^2 = \sum_{i=1}^{n} \frac{\left(Z_i - \bar{Z}\right)^2}{n-1}$
  - T statistic: $\frac{\bar{Z}-H_a}{s/\sqrt{n}}$, where H_a is the hypothesis (taken to be 0 for the null hypothesis).
- The spirit of the T interval assumptions are violated for skewed distributions.
- Other intervals are more useful for highly discrete data.
- Independent group T confidence intervals - comparing means between two different groups in a randomized trial.
  - Can't use a paired T test because the groups are independent and may have different sample sizes.
  - Standard confidence interval to use in this situation: $\bar{Y} - \bar{X} \pm t_{n_x+n_y-2,1-\alpha/2} S_p \left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}$
    - $t_{n_x+n_y-2,1-\alpha/2}$ is the relevant T quantile, where n_i is the number of observations in group i.
    - $\left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}$ is the standard error of the difference. Gets smaller as we collect more data.
    - S_p^2 is the "pooled variance." $S_p^2 = \left[(n_x-1)S_x^2 + (n_y-1)S_y^2\right]/(n_x+n_y-2)$
  - This interval assumes a constant variance across the two groups!
  - If there is some doubt, there is a method for using a different variance per group, which we will discuss later.

- Unequal variances: $\bar{Y} - \bar{X} \pm t_{df} \left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y} \right)^{1/2}$
  - In this case, the distribution does not actually follow a T distribution!
  - We can approximate it with a T distribution with degrees of freedom $df = \frac{\left(S_x^2/n_x + S_y^2/n_y\right)^2}{\left(\frac{S_x^2}{n_x}\right)^2/(n_x-1) + \left(\frac{S_y^2}{n_y}\right)^2/(n_y-1)}$
    - Degrees of freedom may be fractional, but it's OK.
  - To do this in R, use `t.test()`, but set `var.equal=FALSE`.
  - T-statistic: $t = \frac{\bar{Y} - \bar{X}}{\sqrt{S_x^2/n_x + S_y^2/n_y}}$
  - Calculate p-value: `pt(t, df=df, lower.tail=FALSE)`.
    - Multiply by 2 if doing a two-sided test.

Hypothesis testing

- Hypothesis testing is concerned with making decisions using data.
- A null hypothesis is specified that represents the status quo, usually labeled H_0.
- The null hypothesis is assumed to be true and statistical evidence is required to reject it in favor or a research or alternative hypothesis, H_a.
- There are four possible outcomes of our statistical decision process:
  - Correctly accept null.
  - Type I error: select H_a when the null hypothesis is true.
  - Correctly reject null.
  - Type II error: select null hypothesis when H_a is true.
  - As the type I error rate increases, the type II error rate decreases, and vice versa.
- Typical way to do decision making: reject the null hypothesis if $\bar{X}$ is larger than some constant C, chosen such that the probability of the type I error is 0.05.
- Two-sided tests
  - Suppose that we would reject the null hypothesis if in fact the mean was too large or too small.
  - We will reject if the test statistic is either too large or too small.
  - Then we want the probability of rejecting the null hypothesis to be 5%, split equally as 2.5% in the upper tail and 2.5% in the lower tail.
  - Thus, we reject if our test statistic if larger than `qt(0.975)` or smaller than `qt(0.025)`.
  - In general, if you fail to reject the one-sided test, you should fail to reject the two-sided test as well.

P-values

- Most common measure of statistical significance.
- Somewhat controversial among statisticians because they are used often and are commonly misinterpreted.
- What is a p-value?
  - Assume that nothing is going on (null hypothesis) - how unusual is the result we got?

- Approach:
  - Define the hypothetical distribution of a statistic when "nothing" is going on (null distribution).
  - Calculate the statistic with the data we have (test statistic).
  - Compare what we calculated to our hypothetical distribution and see if the value is "extreme" (p-value).
- Formal definition: probability under the null hypothesis of obtaining evidence as extreme or more extreme than that obtained.
  - If the p-value is small, then either H_0 is true and we have observed a rare event, or H_0 is false.
- Example: getting a T statistic of 2.5 for 15 degrees of freedom.
  - What's the probability of getting a T statistic as large as 2.5?
    - `pt(2.5, 15, lower.tail=FALSE)` gives 0.01225.
- Can also think of the p-value as the "attained significance level."
  - Smallest value of α for which we would reject the null hypothesis.
- By reporting a p-value, the reader can perform the hypothesis test at whatever α level they want.
  - If the p-value is less than α, you reject the null hypothesis.
  - For a two-sided hypothesis test, double the smaller of the two one-sided hypothesis test p-values.

Power

- Power is the probability of rejecting the null hypothesis when it is false. Power is a good thing!
  - Comes into play more when you fail to reject the null hypothesis.
- A type II error is failing to reject the null hypothesis when it is false. The probability of a type II error is usually called β, and power is calculated as $Power = 1 - \beta$.
- Example:
  - $H_0 : \mu = 30$ vs. $H_a : \mu > 30$.
  - T-statistic: $\frac{\bar{X}-30}{s/\sqrt{n}}$
  - Power: $P\left(\frac{\bar{X}-30}{s/\sqrt{n}} > t_{1-\alpha,n-1}; \mu = \mu_a\right)$
    - This is equal to α for $\mu_a = 30$ (null hypothesis).
    - It's not equal to α for $\mu_a > 30$, but it approaches α as μ_a goes to 30.
  - We assume the statistic follows a t-distribution under the null hypothesis.
- If we calculate a power of 0.64 for a mean of 32 when the null hypothesis is 30, that means we have a 64% chance of detecting a mean as large as 32.
- Calculating power for Gaussian data
  - We reject if $\frac{\bar{X}-30}{\sigma/\sqrt{n}}h > z_{1-\alpha}$.
    - Equivalently, we reject if $\bar{X} > 30 + Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$
  - Under H_0: $\bar{X} \sim N\left(\mu_0, \sigma^2/n\right)$

- Under H_a: $\bar{X} \sim N\left(\mu_a, \sigma^2/n\right)$
- R code:

```
z <- qnorm(1-alpha)
pnorm(mu0 + z*sigma/sqrt(n), mean=mua, sd=sigma/sqrt(n), lower.tail=false)
```

- When testing the alternative hypothesis, notice that if power is $1-\beta$, then $1-\beta = P\left(\bar{X} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}; \mu = \mu_a\right)$
  - where $\bar{X} \sim N\left(\mu_a, \sigma^2/n\right)$.
  - Unknowns: μ_a, σ, n, β.
  - Knowns: μ_0, α.
  - Specify any 3 of the unknowns and you can solve for the remaining one.
- Other notes
  - The calculation for $H_a : \mu < \mu_0$ is similar to what we've already done.
  - For $H_a : \mu \neq \mu_0$, calculate the one-sided power using $\alpha/2$ (this is only approximately right, it excludes the possibility of getting a large t-statistic in the opposite direction of the truth, but this is only meaningful if μ_0 and μ_a are close to each other).
  - Power goes up as α gets larger.
  - Power of a one-sided test is greater than the power of the associated two-sided test.
  - Power goes up as μ_a gets further away from μ_0.
  - Power goes up as n goes up.
  - Power doesn't need μ_a, σ, n; instead it only needs $\frac{\sqrt{n}(\mu_a-\mu_0)}{\sigma}$
    - The quantity $\frac{\mu_a-\mu_0}{\sigma}$ is called the effect size, the difference in the means in units of standard deviation.
- T-test power
  - The power is $P\left(\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{1-\alpha,n-1}; \mu = \mu_a\right)$.
  - Calculating this requires the non-central t-distribution.
  - `power.t.test` does this very well.
    - If you omit one of the arguments, it will solve for it.
- T-test power example
  - Calculating power:

```
power.t.test(n=16, delta=2/4, sd=1, type="one.sample", alt="one.sided")$power
power.t.test(n=16, delta=2, sd=4, type="one.sample", alt="one.sided")$power
power.t.text(n=16, delta=100, sd=200, type="one.sample", alt="one.sided")$power
## all examples give the same result
## this is because they all have the same ratio of delta to sd.
```

  - Calculating sample size when we give power:

```
power.t.test(power=0.8, delta=2/4, sd=1, type="one.sample", alt="one.sided")$n
```

Multiple testing

- Key ideas:
  - Hypothesis testing and significance analysis are commonly overuse.
  - Correcting for multiple testing avoids false positives or discoveries.
    - Example: do two tests for p-values on the same experiment but only report the smallest.
  - Two key components:
    - Error measure
    - Correction
- Main reason for multiple testing: lots of data!
- Why correct for multiple tests?
  - A p-value of 0.05 doesn't mean much if you did 20 different tests!
- Types of errors
  - Suppose you are testing a hypothesis that a parameter β equals zero versus the alternative (not equal to zero).
  - The possible outcomes are:
	$\beta = 0$	$\beta \neq 0$	Hypotheses
Claim $\beta = 0$	U	T	$m - R$
Claim $\beta \neq 0$	V	S	R
Claims	m_0	$m - m_0$	m
  - Type I error (false positive, V): say that the parameter is not equal to zero when it really is (false alarm rate).
  - Type II error (false negative, T): say that the parameter is zero when it isn't (false dismissal rate).
- Error rates:
  - False positive rate: the rate at which false results ($\beta = 0$) are called significant. $E\left[\frac{V}{m_0}\right]$
    - This is closely related to the type I error rate.
  - Family-wise error rate (FWER): probability of at least one false positive. $Pr\left(V \geq 1\right)$
  - False discovery rate (FDR): the rate at which claims of significance are false. $E\left[\frac{V}{R}\right]$
- Controlling the false positive rate
  - If p-values are correctly calculated, calling all $P < \alpha$ significant will control the false positive rate at level α on average.
  - Problem: suppose that you perform 10000 tests and $\beta = 0$ for all of them.
    - If you call $P < 0.05$ significant, the expected number of false positives is 500.
    - How do we avoid so many false positives?
- Controlling the family-wise error rate
  - The Bonferroni correction is the oldest multiple testing correction.
  - Basic idea:
    - Suppose you do m tests.
    - You want to control FWER at level α such that $Pr\left(V \geq 1\right) < \alpha$.

    - Calculate p-values normally.
    - Set $\alpha_{FWER} = \alpha/m$ and call all p-values less than α_{FWER} significant.
  - Pros: easy to calculate, conservative.
  - Cons: may be very conservative!

- Controlling the false discovery rate
  - This is the most popular correction when performing lots of tests.
  - Basic idea:
    - Suppose you do m tests.
    - You want to control FDR at level α so $E\left[\frac{V}{R}\right]$
    - Calculate p-values normally.
    - Order the p-values from smallest to largest.
    - Call any $p_i \leq \alpha\frac{i}{m}$ significant.
  - Pros: easy to calculate, less conservative.
  - Cons: allows for more false positives, may behave strangely under dependence.

- Adjusted p-values
  - One approach is to adjust the threshold α, but another approach is to calculate "adjusted p-values."
    - In this case, they are not technically p-values any more, so they don't have the same properties of classically defined p-values.
    - But they can be used to control error parameters directly without adjust α.
  - Example:
    - Suppose p-values are $p_1, ..., p_m$.
    - You could adjust them by taking $p_i^{FWER} = max\left(m \cdot p_i, 1\right)$, for each p-value.
    - Then if you call all $p_i^{FWER} < \alpha$ significant, you will control the FWER.
  - R examples:
    - `p.adjust(pValues, method="Bonferroni")`
    - `p.adjust(pValues, method="BH")`

- Note: if there is strong dependence between different tests, there may be problems.
  - Can try `method="BY"` if this is the case.

Bootstrapping

- The bootstrap is a tremendously useful tool for constructing confidence intervals and calculating standard errors for difficult statistics.
- Example: how would you derive a confidence interval for the median?
  - Can do complicated math, but a bootstrap is an easier solution.
- Example: rolling a dice 50 times and calculating the average roll.
  - What if we only have one sample (of 50 rolls)?
  - How can we get a distribution of averages for 50 rolls if we only have one sample of 50 rolls?
  - Bootstrapping says that we should take our one sample of 50 rolls and use the distribution of single dice rolls to generate a population of averages of 50 rolls.

    * Using our observed data to construct an **estimated** population distribution.
    * Using that population distribution to simulate the statistic we are interested in.

- The bootstrap principle
  - If you have a statistic that estimates some population parameter, but don't know its sampling distribution, then you can use the distribution defined by the data to approximate its sampling distribution.
- The bootstrap in practice:
  - Always carried out using simulation.
  - Procedure:
    * Simulate complete data sets from the observed data (with replacement).
      · This is approximately drawing from the sampling distribution of that statistic, at least as far as the data is able to approximate the true population distribution.
    * Calculate the statistic for each simulated data set.
    * Use the simulated statistics to either define a confidence interval or take the standard deviation to calculate a standard error.
- Example: calculating confidence interval for the median of a data set of n observations.
  - Sample n observations **with replacement** from the observed data resulting in one simulated complete data set.
  - Take the median of the simulated data set.
  - Repeat these steps B times, resulting in B simulated medians. (B should be large.)
  - These medians are approximately drawn from the sampling distribution of the median of n observations; therefore we can:
    * Make a histogram of them.
    * Calculate their standard deviation to estimate the standard error of the median.
    * Take the 2.5th and 97.5th percentiles as a confidence interval for the median.
- Example code:

```
library(UsingR)
data(father.son)
x <- father.son$sheight
n <- length(x)
B <- 10000
## Make a matrix where each row is a sample with n observations.
resamples <- matrix(sample(x, n*B, replace=TRUE), B, n)
## Take the median of each row.
medians <- apply(resamples, 1, median)
## Estimated standard error on the median.
sd(medians)
## Estimate a confidence interval for the median.
quantile(medians, c(0.025, 0.975))
```

- The bootstrap is non-parametric: it makes no assumptions about the probability distributions of the variables being assessed.
- Better percentile bootstrap confidence intervals correct for bias.
  - Use the "BCA" interval instead. (???)

Permutation tests

- Used for group comparisons.
- Permutation tests are very powerful and there are several variations:
  - Rank sum test, Fisher's exact test, etc.
- Permutation tests work very well in multivariate settings.
- Example: consider comparing two independent groups using InsectSprays data set.
  - Consider the null hypothesis - that the distribution of the observations from each group is the same.
  - Consider a data frame with count and spray.
  - Permute the spray (group) labels and recalculate the statistic.
    - Mean difference in counts.
    - Geometric means.
    - T-statistic.
  - Calculate the percentage of simulations where the simulated statistic was more extreme (toward the alternative) than the observed.
  - This yields a permutation-based p-value.
- Code example:

```
subdata <- InsectSprays[InsectSprays$spray %in% c("B","C"),]
y <- subdata$count
group <- as.character(subdata$spray)
testStat <- function(w, g) mean(w[g=="B"]) - mean(w[g=="C"])
observedStat <- testStat(y, group)
permutations <- sapply(1:10000, function(i) testStat(y, sample(group)))
observedStat
## [1] 13.25
mean(permutations > observedStat)
## [1] 0
## Using 10000 permutations, we couldn't find a reconfiguration of the group labels
## that led to such an extreme difference.
```