Statistical Inference Writeup

jongman@gmail.com

January 19, 2015

This is a personal writeup of *Statistical Inference* (Casella and Berger, 2nd ed.). The purpose of this note is to keep a log of my impressions during the reading process, so I cannot guarantee the correctness of contents. :-)

Contents

1 Probability Theory

We will not discuss elementary set theory and probability theory.

- Kolmogorov's axiom set defines a probability function for a given σ-algebra.
- Bonferroni's inequality is useful for having a gut estimation of lower bound for events that are hard or impossible to calculate probabilities for. It is equivalent to Boole's inequality.
  - Is a generalization of: $P(A \cap B) \geq P(A) + P(B) - 1$
    - Proof: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Rearranging, we get $P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$.
- Random variables: function mapping from sample space to real numbers.

2 Transformations and Expectations

2.1 Transformations of random variables

This section covers functions of random variables and how to derive their cdf from the cdf of the original RV. Say we have a random variable X with a pdf f_X or a cdf F_X. What is the distribution of $Y = g(X)$?

2.1.1 Monotonic functions

We can only do this for monotonic functions, or at least piecewise monotonic functions. So what do we do when function $Y = g(X)$ is monotonic?

- If monotonically *increasing*,

$$F_Y(y) = P(Y \leq y) = P(g(x) \leq y) = P\left(x \leq g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right)$$

- If decreasing

$$F_Y(y) = P(Y \leq y) = P(g(x) \leq y) = P\left(x \geq g^{-1}(y)\right) = 1 - F_X\left(g^{-1}(y)\right)$$

What do you do when you have pdf, not cdf of the original RV? You don't have to go through the above way; you can differentiate above using chain rule – the below formula takes care of both increasing and decreasing functions.

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|$$

In fact, this is used far more often than the cdf case.

2.1.2 Piecewise monotonic functions

You can also do this when $g()$ is piecewise monotonic. it means, you should be able to partition χ, the domain of the original RV, into contiguous sets so the function is monotonic in each partition. For example, if $f(X) = X^2$, we should split the real line into $(-\infty, 0]$ and $(0, \infty)$.

Let us partition χinto multiple subsets $A_1, A_2, \cdots$ and let $g_1^{-1}, g_2^{-1}, \cdots$ be the inverse of $g()$ in each of the intervals. Then we state without proof

$$f_Y(y) = \sum\left(f_X\left(g_i^{-1}(y)\left|\frac{d}{dy}g_i^{-1}(y)\right|\right)\right)$$

which basically sums up the previous formula for each interval.

2.1.3 Probablity integral transformation

If X has a cdf F_X, and we say $Y = F_X(X)$, then Y is uniformly distributed on $(0, 1)$. This can be understood intuitively: let $F_X(x) = y$. Then $P(Y \leq y) = P(X \leq x) = F_X(x) = y$. This of course assumes monotonicity on F_X's part, which is not always true, but this can be treated technically.

2.2 Expected Values

This section discusses the definition of expected values and their properties. The **linearity of expectation** arises from integration which is how EVs are defined: $\mathrm{E}(A + B) = \mathrm{E}A + \mathrm{E}B$

regardless if A and B are independent or not.

When you want to make transformation of a random variable and take its expected value, you can calculate EVs directly from definition. Otherwise, you can transform the pdf using the above strategy and go from there.

2.3 Moments and Moment Generating Functions

Definitions:

- The nth moment of X, $\mu_n^{'}$ is defined by $\mathrm{E}\left(X^n\right)$.
- The nth central moment of X, μ_n is defined by $\mathrm{E}\left((X-\mu)^n\right)$.
- A moment generating function of X is defined as $M_X\left(t\right) = \mathrm{E}e^{tX} = \int_x e^{tx} f_X\left(x\right) dx$.
- Variance Var$X$ is defined as the second central moment. Of couse, we have the following equality as well:
$$\mathrm{Var}X = \mathrm{E}X^2 - \left(\mathrm{E}X\right)^2$$

A moment generating function can be used to generate moments: we have

$$\frac{d^n}{dt^n} M_X\left(0\right) = EX^n$$

Moment generating functions can be used to identify distributions; if two distributions have same mgf and all moments exist, they are the same distribution. We discuss some theorems regarding convergence of mgfs to another known mgf to prove the convergence of the distribution. It looks like it has more of a theoretical importance, rather than practical.

3 Common Families of Distribution

3.1 Discrete Distributions

Discrete Uniform Distribution The simplest of sorts.

- $P\left(X=x|N\right) = \frac{1}{N}$, $x \in 1, 2, 3, \cdots, N$
- $\mathrm{E}X = \frac{N+1}{2}$, Var$X = \frac{(N+1)(N-1)}{12}$

Hypergeometric Distribution Say we have a population of size N, of which M has a desired property. We take a sample of size K (typically $K \ll M$). What is the probablity that x of those have this property?

- pmf is derived from counting principles, EV and Var from a similar manner to binomial distribution: rewrite the sum as a sum of hypergeometric pmf for a smaller parameter set which equals 1.
- $\mathrm{P}\left(X=x|N,M,K\right) = \binom{M}{x}\binom{N-M}{K-x} / \binom{N}{K}$

- E$X = \frac{KM}{N}$, Var$X = \frac{KM}{N}\left(\frac{(N-M)(N-K)}{N(N-1)}\right)$

Binomial Distribution

- P$\left(X = x|N, p\right) = \binom{N}{x} p^x \left(1-p\right)^{N-x}$
- E$X = np$, Var$X = np\left(1-p\right)$

Poisson Distribution

- P$\left(X = x|\lambda\right) = \frac{e^{-\lambda}\lambda^x}{x!}$
- E$X =$ Var$X = \lambda$
- When a binomial distribution's p is very small, the distribution can reasonably approximated by a Poisson distribution with $\lambda = np$. (Proof uses mgfs)

Negative Binomial Distribution If we want to know the number of Bernoulli trials required to get r successes? Put another way, we are interested in the number of failures before rth success.

- P$\left(Y = y\right) = (-1)^y \binom{-r}{y} p^r \left(1-p\right)^y = \binom{r+y-1}{y} p^r \left(1-p\right)^y$
- E$Y = r\frac{1-p}{p}$ (a simple proof: you can find expected number of failures before each success andsum them, because linearity. Woohoo!)
- Var$Y = \frac{r(1-p)}{p^2} = \mu + \frac{1}{r}\mu^2$
- Negative binomial family includes Poisson as a limiting case (Poisson is also related with Binomial, so this seems natural) but doesn't seem to have a large practical significance.

Geometric Distribution A special case of negative binomial distribution with $r = 1$.

- P$\left(X = x\right) = p\left(1-p\right)^{x-1}$
- E$X = \frac{1}{p}$, Var$X = \frac{1-p}{p^2}$
- "Memoryless" property: history so far has no influence on what will happen from now on. So: P$\left(X > s|X > t\right) = P\left(X > s - t\right)$

3.2 Continuous Distributions

Uniform

Gamma A highly generic & versatile family of distributions.

- A primer on gamma function Γ:
  - $\Gamma\left(\alpha\right) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ which also has a closed form when $\alpha \in$ **N**.
  - Also, the gamma function serves as a generalization of factorial, because $\Gamma\left(\alpha+1\right) = \alpha\Gamma\left(\alpha\right)$ if $\alpha > 0$. (This makes Gamma function easy to evaluate when we know its values between 0 and 1)

  - We have $\Gamma(\alpha) = (\alpha - 1)!$ for integers.

- The full gamma family has two parameters, α(shape parameter, defines peakedness) β the scale parameter (determines spreadness).
- The pdf is set as

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}$$

- E$X = \alpha\beta$, Var$X = \alpha\beta^3$
- Being a generic distribution, it is related to many other distributions. Specifically, many distributions are special cases of Gamma.
  - P$(X \leq x) =$ P$(Y \geq \alpha)$ where $Y \sim$ Poisson(x/β)
  - When $\alpha = p/2$ (p is an integer) and $\beta = 2$, it becomes a chi-squared pdf with p degrees of freedom.
  - When we set $\alpha = 1$, it becomes the exponential distribution pdf with scale parameter β. (The exponential distribution is a continuous cousin of the geometric distribution.)
  - It is also related to the Weibull distribution which is useful for analyzing failure time data and modeling hazard functions.
- Applications: mostly relted with lifetime testing, etc.

Normal without doubt, the most important distribution.

- pdf

$$f\left(X = x|\mu, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

- Proving the pdf integrates to 1 is kind of clunky.
- E$X = \mu$, Var$X = \sigma^2$.
- 68% - 95% - 99% rule.
- Thanks to CLT, normal distribution pops up everywhere. One example would be using normal to approximate binomials. If p is not extreme and n is large, normal gives a good approximation.
  - *Continuity correction*: Say $Y \sim$ Binomial(n, p) and $X \sim$ n$(np, n(1-p)p)$, we can say P$(a \leq Y \leq b) \approx$ P$(a \leq X \leq b)$. However, P$\left(a - \frac{1}{2} \leq X \leq b + \frac{1}{2}\right)$ is a much better approximation, which is clear from a graphical representation.

Beta One of rare distributions which have a domain of $[0, 1]$. Can be used to model proportions.

- Given that $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx$,

$$f(X = x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{a-1}(1-x)^{\beta-1}$$

- Since the Beta function is related to Gamma function $\left(B\left(\alpha,\beta\right)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right)$, it is related to Gamma function as well.
- It can have varying shapes depending on the parameters; unimodal, monotonic, u-shaped, uniform, etc.
- $\mathrm{E}X=\frac{\alpha}{\alpha+\beta}$, $\mathrm{Var}X=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Cauchy A symmetric, bell-shaped curve with undefined expected value. Therefore, it is mostly used as an extreme case against which we test conjectures. However, it pops up in unexpected circumstances. For example, the ratio of two normal variables follows the Cauchy distribution.

$$f\left(x|\theta\right)=\frac{1}{\pi\left(1+\left(x-\theta\right)^2\right)}$$

Lognormal log of which is the normal distribution. Looks like Gamma function.

- $f\left(x|\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi}\sigma x}e^{-(\lg x-\mu)^2/(2\sigma^2)}$
- $\mathrm{E}X=e^{\mu+\left(\sigma^2/2\right)}$, $\mathrm{Var}X=e^{2\left(\mu+\sigma^2\right)}-e^{2\mu+\sigma^2}$

Double Exponential formed by reflecting the exponential distribution around its mean.

3.3 Exponential Families of Distribution

A family of pdfs or pmfs is an exponential family if it can be expressed as:

$$f\left(x|\theta\right)=h\left(x\right)c\left(\theta\right)\exp\left(\sum_i w_i\left(\theta\right)t_i\left(x\right)\right)$$

Many common families are exponential; normal, gamma, beta, binomial, Poisson, etc. Such form has some algebraic advantages: there are equalities which provide a shortcut for calculating the first two moments (EV and Var) with differentiation instead of summation/integration! Will not discuss the actual formulas because they are kind of messy.

The section also discusses natural parameter set of a exponential family: in the above definition, Also, the definition of full and curved exponential families are introduced; not sure about practical significance.

3.4 Scaling and Location Families

If $f\left(x\right)$ is a pdf, then the following function is a pdf as well.

$$g\left(x|\mu,\sigma\right)=\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$$

where μ is the location parameter, σ is the scaling parameter. Expected value will get translated accordingly, and variance will grow by σ^2.

3.5 Inequalities and Identities

Chebychev Inequality if $g(x)$ is a nonnegative function,

$$\mathrm{P}(g(x) \geq r) \leq \frac{\mathrm{E}g(X)}{r}$$

which looks kind of arbitrary (the units do not match) but is useful sometimes. It sometimes provides useful boundaries.

For example, let $g(x) = \left(\frac{x-\mu}{\sigma}\right)^2$ which is the number of standard deviations squared (square is there to make g nonnegative). Letting $r = 2^2$ yields

$$\mathrm{P}\left(\left(\frac{x-\mu}{\sigma}\right)^2 \geq 4\right) \leq \frac{1}{4}\mathrm{E}\frac{(x-\mu)^2}{\sigma^2} = \frac{1}{4}$$

which is a lower bound of elements 2 standard deviations! This did not make any assumption on the distribution of X.

Stein's Lemma with $X \sim \mathrm{n}(\theta, \sigma^2)$ and g that is differentiable and satisfies $\mathrm{E}|g'(X)| < \infty$, $\mathrm{E}[g(X)(X-\theta)] = \sigma^2 \mathrm{E}g'(X)$. Seems obscure, but useful for calculating higher order moments. Also, wikipedia notes that it is useful in MPT.

4 Multiple Random Variables

4.1 Joint and Marginal Distributions

Mostly trivial stuff. We extend the notion of pmf/pdfs by adding more variables, which give us joint pdf/pmfs.

Marginal distributions (distribution of a subset of variables without referencing the other variables) are introduced. Intuitively, they are "compressed" versions of joint pdf/pmfs by integrating them along a subset of parameters.

4.2 Conditional Distributions and Independence

Intuitively, conditional distributions are sliced versions of joint pdf/pmfs. Some of the random variables are observed; what are the distribution of the remaining variables given the observation?

The derivation of conditional pmf is straightforward for discrete RVs, it is the ratio of joint pmf and marginal pmf. This relationship, somewhat surprisingly, holds true for continuous RVs as well.

Two variables X and Y are said to be **independent** when $f(x,y) = f_X(x) f_Y(y)$. Actually, the converse is true as well; if the joint pdf can be decomposed into a product of two functions, one on x and one on y, they are independent.

Consequences of independence:

- $\mathrm{E}\left(g\left(X\right)h\left(Y\right)\right)=\mathrm{E}g\left(X\right)\mathrm{E}h\left(Y\right)$
- Covariance is 0
- You can get the mgf of their sum by multiplying individual mgfs. This can be used to derive the formula for adding two normal variables.

4.3 Bivariate Transformations

This section mainly discusses strategies of taking transformations (sum, product, division) of two random variables. This is analogous to section 2.1 where we discussed transformations of a single variable.

Problem: You have a random vector (X,Y) and want to know about $U=f\left(X,Y\right)$.

Strategy: We have a recipe for transforming a bivariate vector into another. So we transform (X,Y) into (U,V) and take the marginal pdf to get the distribution of U. V is chosen so it will be easy to back out X and Y from U and V; which is essential in the below recipe.

Recipe: Basically, similar to transformation recipe in 2.1. However, the derivative of the inverse function is replaced by *the Jacobian of the transformation*; which is defined as the determinant of the matrix of partial derivatives.

$$J=\frac{\partial x}{\partial u}\frac{\partial y}{\partial v}-\frac{\partial x}{\partial v}\frac{\partial y}{\partial u}$$

Given this, we have

$$f_{U,V}\left(u,v\right)=f_{X,Y}\left(h_1\left(u,v\right),h_2\left(u,v\right)\right)\left|J\right|$$

where $g_1\left(x,y\right)=u$, $h_1\left(u,v\right)=x$, $g_2\left(x,y\right)=y$, $h_2\left(u,v\right)=y$.

Similar to formula in 2.1, this assumes the transformation is 1-to-1, and thus the inverse exists. When this assumption breaks, we can use the same trick as in 2.1 by breaking the domain into sets where in each set the transformation is 1:1.

$$f_{U,V}\left(u,v\right)=\sum_{i=1}^{k}f_{X,Y}\left(h_{1i}\left(u,v\right),h_{2i}\left(u,v\right)\left|J_i\right|\right)$$

which is the formula in 4.3.6.

4.4 Hierarchical Models and Mixture Distributions

Hierarchical models arise when we model a distribution where a parameter has its own distribution. This is sometimes useful in gaining a deeper undestanding of how things work.

As an example, say an insect lays a large number of eggs following Poisson, and individual egg's survival is a Bernoulli trial. Then, the expected number of surviving insect is $X|Y\sim$ binomial$\left(Y,p\right)$ and $Y\sim$ Poisson$\left(\lambda\right)$.

This section also introduces a trivial, but useful equality: $\mathrm{E}X = \mathrm{E}\left(\mathrm{E}\left(X|Y\right)\right)$ This is very intuitive if you think about it, but realizing this makes calculations very easy sometimes. A noncentral chi-squared distribution is given as an example.

A formula for variance in hierarchical model is given:

$$\mathrm{Var}X = \mathrm{E}\left(\mathrm{Var}\left(X|Y\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(X|Y\right)\right)$$

4.5 Covariance and Correlation

This section introduces covariances and correlations. Important stuff, but it only gives rather a basic treatment.

Definitions and identities:

- $\mathrm{Cov}\left(X, Y\right) = \mathrm{E}\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right] = \mathrm{E}XY - \mu_X \mu_Y$
- $\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$
- $\mathrm{Var}\left(aX + bY\right) = a^2 \mathrm{Var}X + b^2 \mathrm{Var}Y + 2ab\mathrm{Cov}\left(X, Y\right)$

This section also introduces bivariate normal distributions.

4.6 Multivariate Distributions

A strategy for transforming a vector of random variable is introduced. Since the Jacobian is well defined for larger matrices, the recipe is more or less the same with the bivariate case.

4.7 Inequalities and Identities.

Analogous to chapter 2, we have a section devoted to inequalities and identities. Apparently, all of these inequalities have many forms, being applied in different contexts. Many of these have popped up in CvxOpt course as well. (Looks like most primitive forms come from field of mathematical analysis.)

Holder's Inequality Let p, q be real positive numbers s.t. $\frac{1}{p} + \frac{1}{q} = 1$ and X, Y are RV. Then we have

$$|\mathrm{E}XY| \leq \mathrm{E}\left|XY\right| \leq \left(\mathrm{E}\left|X\right|^p\right)^{1/p} \left(\mathrm{E}\left|Y\right|^q\right)^{1/q}$$

Cauchy-Schwarz Inequality A special case of Holder's inequality when $p = q = 2$. Then,

$$|\mathrm{E}XY| \leq \mathrm{E}\left|XY\right| \leq \sqrt{\mathrm{E}\left|X\right|^2 \mathrm{E}\left|Y\right|^2}$$

In vector terms, it means $\mathbf{x} \cdot \mathbf{y} \leq |\mathbf{x}|\,|\mathbf{y}|$. This is intuitive, as taking inner product gives chance of things canceling out each other, just like triangular inequality.

Also notable is that this can be used to prove the range of the correlation; just take $|\mathrm{E}\left(X-\mu_X\right)\left(Y-\mu_Y\right)|$ and apply CS inequality. Squaring each side gives

$$\left(\mathrm{Cov}\left(X,Y\right)\right)^2 \leq \sigma_X^2\sigma_Y^2$$

Minkowski's Inequality This feels like an additive version of Holder's Inequality.

$$\left[\mathrm{E}\left|X+Y\right|^p\right]^{1/p} \leq \left[\mathrm{E}\left|X\right|^p\right]^{1/p} + \left[\mathrm{E}\left|Y\right|^p\right]^{1/p}$$

Jensen's Inequality on convex function. Given a convex function g,

$$\mathrm{E}g\left(X\right) \geq g\left(\mathrm{E}X\right)$$

5 Properties of a Random Sample

This chapter deals with several things:

- Definition of random sample
- Distribution of functions of random sample (statistics) and how they converge as we increase sample size
- And of course, the LLN and the CLT.
- Generating random samples.

5.1 Basic Concepts of Random Samples

A random sample $X_1,\cdots,X_n$ is a set of *independent and identically distributed* (iid) RVs. This means we are sampling with replacement from an infinite population. This assumption doesn't always hold, but is a good approximation in a lot of cases.

5.2 Sums of Random Variables from a Random Sample

Sums of random variables can be calculated, of course, using the transformation strategies from Chapter 4. However, since each random variable is iid, the calculation can be simplified greatly.

Also, a definition: Given a vector-or-scalar valued *statistic* $Y=T\left(X_1,X_{2,}\cdots\right)$, the distribution of Y is called the *sampling distribution* of Y.

5.2.1 Basic Statistics

Two most basic statistics are introduced.

Sample mean $\bar{X}$ is defined as

$$\bar{X} = \frac{1}{n} \sum_i X_i$$

Sample variance S^2 is defined as

$$S^2 = \frac{1}{n-1} \sum_i \left(X_i - \bar{X}\right)$$

The formula for sample variance raises some questions; why $n-1$, not n? This was chosen so that $\mathrm{E}S^2 = \sigma^2$. If we were to use n, S^2 would be a biased towards 0.

5.2.2 Bessel's Correction

Using $n-1$ instead of n as the denominator in sample variance is called Bessel's Correction. The legitimacy of this can be proven by taking $\mathrm{E}S^2$ and seeing it equals to σ^2, but the wikipedia page offers some explanation.

By taking sample mean, we are minimizing the squared error of the samples. So unless the sample mean is equal to population mean, sample variance must be smaller than variance measured using the population mean.

Put more intuitively, the sample mean skews towards whatever sample observed, so the variance will be smaller than real.

Another subtle, but *astounding* point mentioned is that by Jensen's inequality, S is a *biased* estimator of standard deviation – it underestimates! Since square root is a concave function, according to Jensen's inequality we have

$$\mathrm{E}\sqrt{S^2} \leq \sqrt{\mathrm{E}S^2} = \sqrt{\sigma^2} = \sigma$$

Also, it is noted that there is no general formula for an unbiased estimate of the standard deviation!

5.2.3 Sampling Distributions of Sample Mean and Sample Variance

- $\mathrm{E}\bar{X} = \mu$: the expected value of sample mean is the population mean. LLN will state that it will converge to the population mean *almost surely* as the sample size grows.
- $\mathrm{Var}\bar{X} = \frac{\sigma^2}{n}$: this follows from that the random variables are independent.

  $$\mathrm{Var}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum X_i\right) = \frac{\mathrm{Var}X_1}{n} = \frac{\sigma^2}{n}$$

  So, the variance decreases linearly with the sample size.

- $\mathrm{E}S^2 = \sigma^2$. This can be shown algebraically.

5.2.4 Using mgfs to Find Sampling Distributions

We can plug the statistic's definition into mgf, and simplify, praying the mgf would be a recognizable form. For example, this is how we show the mean of iid normal variables $X_i \sim \mathrm{n}\left(\mu, \sigma^2\right)$ follows $\mathrm{n}\left(\mu, \frac{\sigma^2}{n}\right)$.

5.3 Sampling From the Normal Distribution

More notable facts when we sample from the normal distribution.

- $\bar{X}$ and S^2 are independent random variables.
- $\bar{X}$ has a $\mathrm{n}\left(\mu, \sigma^2/n\right)$ distribution (proven by mgfs as noted above)
- $(n-1)\,S^2/\sigma^2$ has a chi-squared distribution with $n-1$ degrees of freedom.

5.3.1 Student's t-distribution

If $X_i \sim \mathrm{n}\left(\mu, \sigma^2\right)$, we know that

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{n}\,(0,1)$$

However, in most cases μ and σ are unknown parameters. This makes it hard to make inferences about one of them. We can approximate σ by S: this gives us another distribution, however it makes it easier to make inferences about μ. So, the statistic

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

is known to follow Student's t-distribution with $n-1$ degrees of freedom. The concrete distribution could be found using the independence between $\bar{X}$ and S , and their respective distributions (normal and chi-squared).

5.3.2 Fisher's F-distribution

Now we are comparing variances between two samples. One way to look at this is the variance ratio. There are two variance ratios; one sample, one population. Of course, we don't know about population variance ratios. However, the ratio between the two ratios (I smell recursion)

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

is known to follow F-distribution. The distribution is found by noting that in the second representation above, both the numerator and the denomiator follows chi-squared distributions.

5.4 Order Statistics

Order statistics of a random sample are the sorted sample values. This seems obscure, but indeed is useful where point estimation depends on the minimum/maximum value observed. The pmf/pdf of those can be derived from noticing that:

- Say we want to find P $\left(X_{(j)} \leq x_i\right)$ ($X_{(j)}$ is the$j$th smallest number).
- Each random variable is now a Bernoulli trial, with $F_X\left(x_i\right)$ probability of being $\leq x_i$.
- So the cumulative functions can be derived from binomial distribution of each variable (for discrete) or something similar (continous).

$$f_{X_{(j)}}\left(x\right) = \frac{n!}{\left(j-1\right)!\left(n-j\right)!} f_X\left(x\right)\left[F_X\left(x\right)\right]^{j-1}\left[1-F_X\left(x\right)\right]^{n-j}$$

5.5 Convergence Concepts

The main section of the chapter. Deals with how sampling distributions of some statistics change when we send the sample size to infinity.

5.5.1 Definitions

Say we have a series of random variables $\{X_i\}$: X_n being the statistic value when we have n variables. We study the behavior of $\lim_{n\to\infty} X_n$.

There are three types of convergences:

Convergence in Probability when for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}\left(\left|X_n - X\right| \geq \epsilon\right) = 0$$

Almost Sure Convergence when

$$\mathrm{P}\left(\lim_{n\to\infty} \left|X_n - X\right| \geq \epsilon\right) = 0$$

which is a much stronger guarantee and implies convergence in probability.

Convergence in Distribution when their mgfs converge and they have the same distribution.

This is equivalent to convergence in probability when the target distribution is a constant.

5.5.2 Law of Large Numbers

Given a random sample $\{X_i\}$ where each RV has finite mean, the sample mean converges almost surely to μ, the population mean.

There is a weak variant of LLN, which states it converges in probability.

5.5.3 Central Limit Theorem

The magical theorem! When we have a sequence of iid RVs $\{X_i\}$ with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 > 0$ then

$$\lim_{n\to\infty} \frac{\left(\bar{X}_n - \mu\right)}{\sigma/\sqrt{n}} \sim \mathrm{n}\left(0,1\right)$$

which also means

$$\lim_{n\to\infty} \bar{X}_n \sim \mathrm{n}\left(\mu, \sigma^2\right)$$

Another useful, and intuitive theorem is mentioned:

Slutsky's Theorem if Y_n converges to a constant a in probability, and X_n converges to X in distribution,

- $X_n Y_n = aX$ in distribution
- $X_n + Y_n = a + X$ in distribution

Slutsky's theorem is used in proving that normal approximation with *estimated variance* goes to standard normal as well. We know

$$\lim_{n\to\infty} \frac{\left(\bar{X}_n - \mu\right)}{\sigma/\sqrt{n}} \sim \mathrm{n}\left(0,1\right)$$

and we can prove

$$\lim_{n\to\infty} \frac{\sigma}{S} = 1$$

in probability. Multiplying those yields

$$\lim_{n\to\infty} \frac{\left(\bar{X}_n - \mu\right)}{S/\sqrt{n}} \sim \mathrm{n}\left(0,1\right)$$

Marvelous!

5.5.4 Delta Method

Delta method is a generalized version of CLT. Multiple versions of it are discussed, however here I will only state the most obvious univariate case.

$$\lim_{n\to\infty} Y_n \sim \mathrm{n}\left(\mu, \frac{\sigma^2}{n}\right) \implies \lim_{n\to\infty} g\left(Y_n\right) \sim \mathrm{n}\left(g\left(\mu\right), \frac{g'\left(\mu\right)^2 \sigma^2}{n}\right)$$

When Y_n converges to μ with normal distribution when we increase n with virtual certainty, a function of Y_n converges to a normal distribution. In the limiting case, both are virtually a constant, so this doesn't surprise me much.

6 Principles of Data Reduction

This chapter seems to be highly theoretic. Without much background, it is hard to say why this material is needed. However, it looks like this chapter is closely related with the next chapter - point estimation. That makes sense, because estimating point parameters are essentially data reduction.

6.1 The Sufficiency Principle

6.1.1 Sufficient Statistic

Colloquially, a sufficient statistic *captures* all information regarding a parameter θ in a sample x; there are no remaining information to be obtained by consulting the actual sample. Formally, a statistic $T(X)$ is a sufficient statistic if the conditional distribution $P_\theta(X|T(X))$ does not depend on θ.

The practical upshot of this is that this justifies only reporting means and standard deviations of a given sample; if we assume the population is normal, they are sufficient statistics which contain all the information we can infer about the population. However, remember that these are model dependent; the population might be coming from a different family with different parameters – those might not be entirely captured by the statistic.

6.1.2 Factorization Theorem

The definition could be directly used to verify if a given statistic is sufficient or not, but practically the following theorem makes it easier to identify sufficient statistics. If the joint pdf $f(x|\theta)$ could be factored as a product of two parts:

$$f(x|\theta) = g(T(x)|\theta) \cdot h(x)$$

where $h(x)$ does not depend on θ, then $T(X)$ is a sufficient statistic. This makes intuitive sense; if T is sufficient, the probability of us seeing x is related with only two things: a function of the statistic (given θ) and a function unrelated to θ. If there were a part with another input involved, we wouldn't be able to make inferences about θ with only T.

6.1.3 Minimal Sufficient Statistics

There are lots of sufficient statistics, some more compact than the other. So what is a minimal sufficient statistic? A sufficient statistic $T(X)$ is minimal when for any other sufficient statistic $T'(X)$, $T(X)$ is a function of $T'(X)$. In other words, take any sufficient statistic, and we can use it to derive a minimal sufficient statistic. For example, the sample mean $\bar{x}$ is a sufficient statistic for population mean μ. On the other hand, the sample itself is a sufficient statistic but not minimal – you can derive $\bar{x}$ from the sample, but not vice versa.

6.1.4 Ancillary Statistics

Ancillary statistic contains no information about θ. Paradoxically, an ancillary statistic, when used in conjunction with other statistics, does contain valuable information about θ. The following is a good example; let X have the following discrete distribution

$$P_\theta\left(X=\theta\right)=P_\theta\left(X=\theta+1\right)=P_\theta\left(X=\theta+2\right)=\frac{1}{3}$$

Now, say we observe a sample and take the *range* of the sample $R=X_{(N)}-X_{(1)}$. It is obvious that the range itself has no information regarding θ. However, say we know the mid-range statistic as well; $M=\left(X_{(1)}+X_{(N)}\right)/2$. The mid-range itself can be used to guess θ, but if you combine it with the range, suddely you can nail the exact θ if $R=2$.

Of course, in the above case, the statistic M was not sufficient. What happens when we have a minimal sufficient statistic? The intuition says they should be independent. However, it could not be the case. In fact, the pair $\left(X_{(1)},X_{(N)}\right)$ is minimal sufficient and R is closely related to them! So they are not independent at all.

6.1.5 Complete Statistics

For many important situations, however, the minimal sufficient statistic is indeed independent of ancillary variables. The notion of complete statistic is introduced, but the definition is very far from intuitive. It goes like this: let $T\left(X\right)\sim f\left(t|\theta\right)$. If $\forall\theta E_\theta g\left(T\right)=0 \implies \forall\theta P_\theta\left(g\left(T\right)=0\right)$, $T(X)$ is called a complete statistic. Colloquially, $T\left(X\right)$ has to be uncorrelated with all unbiased estimators of 0.

So WTF does this mean at all? It looks like we will get more context in the next chapter. The most intuitive explanation I could get was from here. It goes like

> Intuitively, if a nontrivial function of T has mean value not dependent on θ, that mean value is not informative about θ and we could get rid of it to obtain a sufficient statistic ”simpler”.

Hopefully reading chapter 7 will give me more background and intuition so I can revisit this.

6.2 The Likelihood Principle

The likelihood function is defined to be

$$L\left(\theta|x\right)=f\left(x|\theta\right)$$

We can use likelihood functions to compare the *plausibility* of various parameter values. Say if we don’t know the parameter and observed x. If for two parameters θ_1 and θ_2, if we have $L\left(\theta_1|x\right)>L\left(\theta_2|x\right)$, we can say θ_1 is more plausible than θ_2. Note we used “plausible” rather than “probable”.

The likelihood principle is an important principle used in later chapters on inferences: if x and y are two sample points such that $L\left(\theta|x\right)$ is proportional to $L\left(\theta|y\right)$, that is, there is a constant

$C(x,y)$ such that $L(\theta|x) = C(x,y)\, L(\theta|y)$, then the conclusions drawn from x and y should be identical.

7 Point Estimation

This chapter deals with estimating parameters of a population by looking at random samples. The first section discusses strategies for finding estimators; the second section deals with ways to evaluate the estimators.

7.1 Methods of Finding Estimators

7.1.1 Method of Moments for Finding Estimator

Basically, you can equate expected values of arbitrary functions of the random sample with realized value to estimate parameters. Method of moments use the first k sample moments to achieve this. Suppose we have $X_i \sim \mathrm{n}\left(\theta, \sigma^2\right)$. The first moment would be θ, the second (uncentered) moment would be $\theta^2 + \sigma^2$. So we have a system of equations

$$\begin{bmatrix} \bar{X} \\ \frac{1}{n}\sum X_i^2 \end{bmatrix} = \begin{bmatrix} \theta \\ \theta^2 + \sigma^2 \end{bmatrix}$$

On the left hand side there are concrete values from the random sample; so solving this is trivial. Also note solving the above for σ gives

$$\tilde{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum \left(X_i - \bar{X}\right)^2$$

which does not have Bessel's correction. This is an example where the method of moments come short; it is a highly general method where you can fall back on, but in general it produces relatively inferior estimators[1].

Interesting Example: Estimating Both Parameters Say we have $X_i \sim \text{binomial}\,(k, p)$ where both k and p are unknown. This might be an unusual setting, but here's an application: if we are modeling crime rates – you don't know the actual number of crimes committed k and the reporting rate p. In this setting, it is hard to come up with any intuitive formulas. So we could use the method of moments as a baseline.

Equating the first two moments, we get:

$$\begin{bmatrix} \bar{X} \\ \frac{1}{n} X_i^2 \end{bmatrix} = \begin{bmatrix} kp \\ kp\,(1-p) + k^2 p^2 \end{bmatrix}$$

Solving this gives us an approximation for k:

[1] However, this is still a maximum-likelihood estimate.

$$\tilde{k} = \frac{\bar{X}^2}{\bar{X} - (1/n) \sum \left(X_i - \bar{X}\right)^2}$$

which is a feat in itself!

7.1.2 Maximum Likelihood Estimation

MLE is, by far, the most popular technique for deriving estimators. There are multiple ways to find MLEs. Classic calculus differentiation is one, exploiting model properties is two, using computer software and maximizing it numerically is three.

Couple things worth noting: using log-likelihood sometimes makes calculations much easier, so it's a good idea to try taking logs. Another thing is that MLE can be unstable - if inputs change slightly, the MLE can change dramatically. (I wonder if there's any popular regularization scheme for MLEs. Maybe you can still use ridge style regularization mixed with cross validation.)

There's a very useful property of maximum likelihood estimators called *the invariance property of maximum likelihood estimators*, which says that if $\hat{\theta}$ is a MLE for θ, then $r\left(\hat{\theta}\right)$ is a MLE for $r(\theta)$. This is very useful to have to say the least.

7.1.3 Bayes Estimators

The Bayesian approach to statistics is fundamentally different from the classical frequentist approach, because they treat parameters to have distribution themselves. In this approach, we have a *prior distribution of* θ which is our belief before seeing the random sample. After we see some samples from the population, our belief is *updated* to the *posterior distribtion* following this formula:

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\,\pi(\theta)}{\int f(\mathbf{x}|\theta)\,\pi(\theta)\,d\theta}$$

which is just the Bayes formula.

Conjugate Families In general, for any sampling distribution, there is a natural family of prior distributions, called the conjugate family. For example, binomial family's conjugate family is the beta family. Let's say if we have a binomial (n, p) distribution where only p is unknown. If our prior distribution of p is beta (α, β), then after observing a sample y our posterior distribution updates as: beta $(y + \alpha, n - y + \beta)$. So classy and elegant! However it is debatable that a conjugate family always should be used.

Also, normals are their own conjugates.

7.1.4 The EM Algorithm

The EM algorithm is useful when the likelihood function cannot be directly maximized; due to missing observations or latent variables which are not observed. The setup goes like this;

given a parameter set θ, there are two random variables $\mathbf{x}$ and $\mathbf{y}$ of which only $\mathbf{y}$ is observed. The marginal distribution for $\mathbf{y}$, $g(\mathbf{y}|\theta)$, is unknown, however we do know $f(\mathbf{y},\mathbf{x}|\theta)$. Since $\mathbf{x}$ is not observed (or incomplete), it is hard to estimate θ from given $\mathbf{y}$.

The EM algorithm solves it by an iterative approach. We start with a guess of the parameter $\theta^{(0)}$, and the following two steps are repeated:

- Expectation step: Given the latest guess for the parameter $\theta^{(i)}$, and the observed $\mathbf{y}$, find the expected $\mathbf{x}$.
- Maximization step: Given the $\mathbf{x}$ and $\mathbf{y}$, find the parameter $\theta^{(i+1)}$ that is most plausible for this pair of observations.

It can be proven that the likelihood from successive iterations of parameters are nondecreasing, and will eventually converge.

EM Algorithm, Formally The above two steps are just colloquial descriptions; the actual expected $\mathbf{x}$ is not directly calculated. In the expectation step, the following *function* is created:

$$Q\left(\theta|\theta^{(t)}\right) = \mathrm{E}_{\mathbf{X}|\mathbf{Y},\theta}\left[\log L\left(\theta;\mathbf{X},\mathbf{Y}\right)\right]$$

Let's see what this function is trying to do; since we have a guess for the parameter $\theta^{(t)}$, we now can plug this into the joint pdf and find the distribution of the pair $(\mathbf{x},\mathbf{y})$. For each concrete pair of observations, we can calculate the log-likelihood of θ using this observation. Q is simply a weighted average of these likelihoods. Q is a function of θ, and answers this question:

> Let's say $(\mathbf{x},\mathbf{y}) \sim f\left(\mathbf{x},\mathbf{y}|\theta^{(t)}\right)$. Now what is the EV of log likelihood for a given θ?

Now the second step puts this function through argmax operator.

7.2 Methods of Evaluating Estimators

This section is much more involved than the previous section, and discusses multiple strategies of evaluating estimators – also discusses some strategies to improve existing estimators.

7.2.1 Mean Squared Error

A simple, intuitive way of measuring the quality of the estimator is taking the MSE:

$$\mathrm{E}_\theta\left(W-\theta\right)^2$$

note this is a function of the parameter θ. Also, MSE has a very intuitive decomposition:

$$\mathrm{E}_\theta\left(W-\theta\right)^2 = \mathrm{Var}_\theta W + \left(\mathrm{E}_\theta W-\theta\right)^2$$

The first term is the variance of the parameter, while the second term is the square of the bias. Variance represents the precision of the estimator; if variance is high, even if the estimator works in average, it cannot be trusted. If the bias is high, it will point in a wrong direction. So this tradeoff is a topic that is discussed throughout this chapter.

Quantifying Bassel's Correction We know $S^2 = \sum \left(X_i - \bar{X}\right)^2 / (n-1)$ is an unbiased estimator of σ^2, thus the bias is 0. However, we also know $\frac{n}{n-1} S^2$ is the maximum likelihood estimator. Which of these are better in terms of MSE? Carrying out the calculations tells you the MLE will give you smaller MSE; so by accepting some bias, we could reduce the variance and the overall MSE. Which estimator we should ultimately use is still debatable.

No Universal Winner Since MSE are functions of parameters, it often is the case the MSE of two estimators are incomparable, because one can outperform another on a certain parameter set, but can underperform on another parameter set. In the text, this is shown with an example of estimating p from binomial (n, p) where n is known. Two approaches are showcased; one is just using the MLE $\sum X_i / n$, another is taking a Bayesian estimate from a constant prior.

If you plot the MSE against different values of p, MSE from the MLE is a quadractic curve; but the MSE from the Bayesian estimate is a horizontal line. Therefore, one doesn't dominate the other – plotting for different n will reveal the Bayesian approach gets better as we increase n.

7.2.2 Best Unbiased Estimator

One way of solving the bias-variance tradeoff problem is to ignore some choices! We can restrict ourselves to unbiased estimators and try to minimize variance. In this setting, the best estimator is called a *uniform minimum variance unbiased estimator* (UMVUE).

Finding UMVUE is a hard problem; it would be hard to prove that the given estimator is the global minimum.

Finding Lower Bounds The *Cramer-Rao Theorem* sometimes helps; it is an application of the Cauchy-Schwartz inequality which provides a lower bound for the variance of an unbiased estimator. So if this limit is reached, we can say for sure we have a UMVUE. This is indeed attainable for some distributions - such as Poisson.

The big honking problem of CR theorem is that the bound is not always *sharp* – sometimes the bound is not attainable. Since CR is an application of CS inequality, we can state the condition under which the equality will hold. This can either give us hints about the shape of the UMVUE, or let us prove it is unreachable.

Sufficiency and Completeness for UMVUE Somewhat surprising and arbitrary - but sufficient or complete statitics can be used to improve unbiased estimators, by conditioning the estimator on the sufficient statistic! The Rao-Blackwell theorem specifies that if W is any unbiased estimator of an arbitrary function of θ, $\tau(\theta)$, the following function is a uniformly ***better*** unbiased estimator:

$$\phi(T) = \mathrm{E}(W|T)$$

This was a WTF moment for me. However, note that

$$\begin{aligned}\mathrm{Var}_\theta W &= \mathrm{Var}_\theta[\mathrm{E}(W|T)] + \mathrm{E}[\mathrm{Var}(W|T)] \\ &\geq \mathrm{Var}_\theta[\mathrm{E}(W|T)]\end{aligned}$$

for any T – you can condition on *any* statistic, even totally irrelevant, and it can never actually hurt. Then why the requirement for sufficiency? If not, the estimator will become a function of θ so it isn't exactly an estimator.

Can we go further than that? Can we create a UMVUE from sufficient statistics? We analyze the properties of a best estimator to find out. Suppose we have an estimator W satisfying $\mathrm{E}_\theta W = \tau(\theta)$. Also say we have an unbiased estimator of 0, U. Then the following estimator

$$\phi_a = W + aU$$

is still an unbiased estimator of $\tau(\theta)$. What is the variance of this?

$$\mathrm{Var}_\theta \phi_a = \mathrm{Var}_\theta(W + aU) = \mathrm{Var}_\theta W + a^2 \mathrm{Var}_\theta U + 2a\mathrm{Cov}_\theta(W, U)$$

Now, we can always concoct a U such that $\mathrm{Cov}_\theta(W, U) < 0$ so $\mathrm{Var}_\theta \phi_a < \mathrm{Var}_\theta W$. Note that U is essentially a random noise for estimation - yet it can give us an improvement. This doesn't make sense at all. Theorem 7.3.20 actually brings order, by stating that W is the UMVUE iff it is uncorrelated with all unbiased estimators of 0.

Wait, that definition sounds familiar. Recall the definition of complete statistics? Now we have to guess they are somewhat related. They indeed are, by Theorem 7.3.23. This states:

> Let T be a complete sufficient statistic for a parameter θ, and let $\phi(T)$ be any estimator based only on T, then $\phi(T)$ is the unique best unbiased estimator of its expected value.

7.2.3 General Loss Functions

MSE is a special case of loss function, so we can generalize on that. Absolute error loss and squared loss are the most common functions, but you can come up with different styles, of course.

Evaluating an estimator on a given loss function $L(\theta, a)$ is done by a risk function:

$$R(\theta, \delta) = \mathrm{E}_\theta L(\theta, \delta(X))$$

which is still a function of the parameter. Also note when we use squared loss function, R becomes MSE.

Stein's Loss Function For scale parameters such as σ^2, the values are bounded below at 0. Therefore, a symmetric loss function such as MSE can be penalizing overestimation because the penalty can grow infinitely only in one direction. An interesting class of loss function, which tries to overcome this problem, is introduced:

$$L\left(\sigma^2, a\right) = \frac{a}{\sigma^2} - 1 - \log \frac{a}{\sigma^2}$$

which goes to infinity as a goes to either 0 or infinity. Say we want to estimate σ^2 using an estimator $\delta_b = bS^2$ for a constant b. Then

$$R\left(\sigma^2, \delta_b\right) = \mathrm{E}\left(\frac{bS^2}{\sigma^2} - 1 - \log \frac{bS^2}{\sigma^2}\right) = b - \log b - 1 - \mathrm{E} \log \frac{S^2}{\sigma^2}$$

which is minimized at $b = 1$.

Bayes Risk Instead of using the risk function as a parameter of θ, we can take a prior distribution on θ and getting the expected value. This takes us to the Bayes risk:

$$\int_\Theta R\left(\theta, \delta\right) \pi\left(\theta\right) d\theta$$

Finding the estimator δ which minimizes the Bayes risk seems daunting but is tractable. Note:

$$\begin{aligned}\int_\Theta R\left(\theta, \delta\right) \pi\left(\theta\right) d\theta &= \int_\Theta \left(\int_\chi L\left(\theta, \delta\left(x\right)\right) f\left(x|\theta\right) dx\right) \pi\left(\theta\right) d\theta \\ &= \int_\chi \left[\int_\Theta L\left(\theta, \delta\left(x\right)\right) \pi\left(\theta|x\right) d\theta\right] m\left(x\right) dx\end{aligned}$$

The quantity in square brackets, which is the expected loss given an observation, is called the posterior expected loss. It is a function of x, not θ. So we can minimize the posterior expected loss for each x and minimize Bayes risk. A general recipe for doing this is not available, but the book contains some examples of doing this.

8 Hypothesis Testing

This chapter deals with ways of testing *hypotheses*, which are statements about the population parameter.

8.1 Terminology

In all testing, we use two hypothesis; the null hypothesis H_0 and the alternative hypothesis H_1, one of which we will accept and the other we will reject. We can always formulate the hypothesis as two sets: H_0 is $\theta \in \Theta_0$ and H_1 is $\theta \in \Theta_0^C$. We decide which hypothesis to accept on the basis of a random sample. The subset of the sample space which will make us reject H_0 is called the *rejection region*.

Typically, a hypothesis test is performed by calculating a *test statistic* and rejecting/accepting H_0 if the statistic falls into a specific set.

8.2 Methods of Finding Tests

We discuss three methods of constructing tests.

8.2.1 Likelihood Ratio Tests

Likelihood Ratio Tests are very widely applicable, and are also optimal in some cases. We can prove t-test is a special case of LRT. The LRT calculates the ratio between the maximum likelihood, and the maximum likelihood given the null hypothesis. The test statistic λ is defined as:

$$\lambda\left(x\right) = \frac{\sup_{\Theta_0} L(\theta|x)}{\sup_{\Theta} L(\theta|x)} = \frac{\sup_{\Theta_0} L(\theta|x)}{L\left(\hat{\theta}|x\right)} \qquad (\hat{\theta} \text{ is MLE of } \theta)$$

When do we reject H_0? We want to reject the null hypothesis when the alternative hypothesis is much more plausible; so the smaller λ is, the more likely we will reject H_0. Therefore, the rejection region is $\{x|\lambda\left(x\right) \leq c\}$ where $0 \leq c \leq 1$.

Also note that λ based on a sufficient statistic of θ is equivalent to regular λ.

Example: Normal LRT for testing $\theta = \theta_0$, known variance

Setup: we are drawing sample from n $(\theta, 1)$ population. H_0: $\theta = \theta_0$ and H_1: $\theta \neq \theta_0$. Since the MLE is $\bar{x}$, the LRT statistic is:

$$\lambda\left(x\right) = \frac{f\left(x|\theta_0\right)}{f\left(x|\bar{x}\right)} = \text{(some simplification)} = \exp\left[-n\left(\bar{x}-\theta_0\right)^2/2\right]$$

Say we reject H_0 when $\lambda \leq c$; we can rewrite this condition as

$$\left|\bar{x}-\theta_0\right| \geq \sqrt{-2\left(\log c\right)/n}$$

Therefore, we are simply testing the difference between the sample mean and the asserted mean with a positive number.

Example: Normal LRT for testing $\mu < \mu_0$, unknown variance

Setup: we are drawing from n $\left(\mu, \sigma^2\right)$ where both parameters are unknown. H_0: $\mu \leq \mu_0$ and H_1: $\mu > \mu_0$. The unrelated parameter σ^2 is called a *nuisance parameter.* It can be shown (from exercise 8.37) that the test relying on this statistic is equivalent to Student's t test.

8.2.2 Bayesian Tests

Bayesian tests will obtain posterior distribution from the prior distribution and the sample, using the Bayesian estimation technique discussed in Chapter 7. And we use the posterior distribution to run the test. We might choose to reject H_0 only if $P\left(\theta \in \Theta_0^C | X\right)$ is greater than some large number, say 0.99.

8.2.3 Union-Intersection and Intersection-Union Tests

Ways of structuring the rejection region when the null hypothesis is not simple. For example, the null hypothesis Θ_0^C is best expressed by an intersection or union of multiple sets. If we have tests for each individual sets that constitute the union/intersection, how can we test the entire hypothesis?

- When the null hypothesis is an intersection of sets, *any* test failing will result in the rejection of the null hypothesis.
- When the null hypothesis is a union of sets, any test passing will let us reject the null hypothesis; null hypothesis is rejected only when all of the tests fail.

8.3 Evaluating Tests

How do we assess the goodness of tests?

8.3.1 Types of Errors and Power Function

Since the null hypothesis could be true or not, and we can either accept or reject it, there are four possibilities in a hypothesis test, which can be summarized in the below table:

		Decision	
		Accept H_0	Reject H_0
Truth	H_0	Correct	Type I Error
	H_1	Type II Error	Correct

To reiterate:

- Type I Error incorrectly rejects H_0 when H_0 is true. Since we usually want to prove the alternative hypothesis, this is a *false positive* error, when we assert H_1 when it is not. Often this is the error people want to avoid more.

- Type II Error incorrectly accepts H_0 when H_1 is true. This is a *false negative* error; the alternative hypothesis is falsely declared negative.

The probabilities of these two errors happening are characterized by a *power function*. A power function takes θ as an input, and calculates the probability that X will be in the rejection region:

$$\beta\left(\theta\right) = P_\theta\left(X \in R\right)$$

What is an ideal power function? We want $\beta\left(\theta\right)$ to be 0 when $\theta \in \Theta_0$, and 1 when $\theta \in \Theta_0^C$. We can plot β with regards to all possible values of θ, and compare the different tests.

Example: Normal Power Function

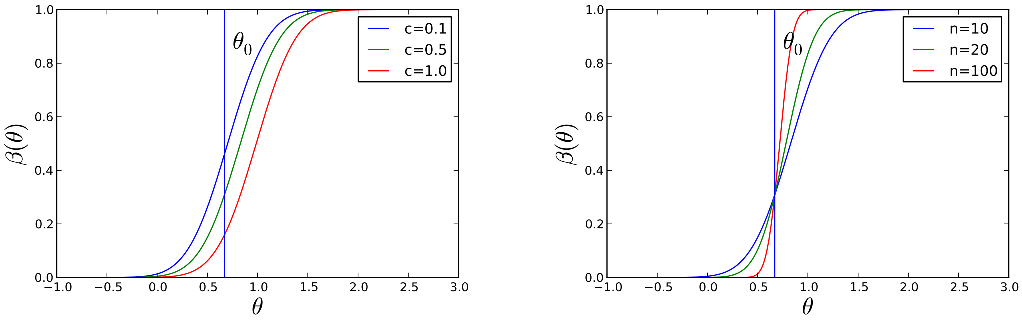
Setup: drawing from n $\left(\theta, \sigma^2\right)$, where σ^2 is known. H_0: $\theta \leq \theta_0$, H_1: $\theta > \theta_0$. Say we have a test that rejects H_0 when

$$\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c$$

The power function of this test is

$$\begin{aligned}
\beta\left(\theta\right) &= P_\theta\left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c\right) \\
&= P_\theta\left(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right) \\
&= P\left(Z > c + \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

where Z is a standard normal random variable. As we increase θ from $-\infty$ to ∞, this probability will go from 0 to 1. Changing c will change when we reject H_0 and affect the error probabilities.



(a)

(b)

The figure (a) gives an example of different c – since H_0 is true when θ is left of θ_0, the type I error probability is depicted by the distance between the x-axis and the power function curve. On the right of θ_0, the type II error probability is depicted by the distance between $y = 1$ and the curve. We can see that increasing c trades off the maximum type I error and the maximum type II error: if we have a higher c, we have a more strict criteria for rejecting H_0. This results in lower false positive, but in higher false negative.

Consequences of the Power Function

Figure (b) above shows the power function for different ns. We see as we increase n, the power function approaches the ideal step function. Typically, the power function will depend on the sample size n. If n can be chosen by the experimenter, considering the power function will be useful to determine the sample size before the experiment is performed.

Size and Level of Tests

For a fixed sample size, it is usually impossible to make both types of error probability arbitrarily small. As most people care more about Type I errors (since alternative hypothesis are what we *want* to prove), we classify tests by the maximum possible type I error across all possible parameters in Θ_0: $\sup_{\theta \in \Theta_0} \beta\left(\theta\right)$

- We want these numbers to be low, since $\beta\left(\theta\right)$ is ideally 0 for Θ_0. Lower size/level means more powerful tests!
- Note as we do not have a prior distribution of parameters, we use the maximum error probability rather than the expected error probability.
- When this bound α is tight ($\sup_{\theta \in \Theta_0} \beta\left(\theta\right) = \alpha$), the test is called a test of size α. When this bound is not tight ($\sup_{\theta \in \Theta_0} \beta\left(\theta\right) \leq \alpha$), the test is called a test of level α.
    - So levels are upper bounds of test powers; every test is a level 1 test but that is not informative.

Unbiased Tests

A test with power function $\beta\left(\theta\right)$ is *unbiased* if

$$\sup_{\theta \in \Theta_0} \beta\left(\theta\right) \leq \inf_{\theta \in \Theta_0^C} \beta\left(\theta\right)$$

Colloquially, if when the alternative hypothesis is true, it should be more likely to reject the null hypothesis than when the null hypothesis is true.

8.3.2 Most Powerful Tests: Uniformly Most Powerful Tests

Depending on the problem, there might be multiple level α tests. What should we use among them? Since we have only controlled for Type I errors so far, it is a good idea to look at Type II errors this time. The *Uniformly Most Powerful* test is the most powerful test across all $\theta \in \Theta_0^C$: its power function β should be larger than any other test's power function, for any $\theta \in \Theta_0^C$.

Neymann-Pearson Lemma

UMP has a very strong argument; we can easily imagine situations where the UMP does not exist. However, in some cases, they do. When you are only considering two hypothesis H_0: $\theta = \theta_1$ and H_1: $\theta = \theta_2$, accepting H_0 when and only when $f(x|\theta_1) < kf(x|\theta_0)$ for some $k \geq 0$ is a UMP level α test, where $\alpha = P_{\theta_0}(X \in R)$.

Colloquially, if your test is formed like above, it is a UMP for its own class (level α).

Extending Neyman-Pearson Lemma To A One-Sided Test

In case of a one-sided hypothesis (for example, $\theta > \theta_0$), we can extend the above lemma to find a UMP.

First, we need to define the following concept: a pdf has a MLR (monotone likelihood ratio) property if: whenever $\theta_2 > \theta_1$, the likelihood ratio $f(t|\theta_2)/f(t|\theta_1)$ is a monotone function of t. (This holds for all the regular exponential families $g(t|\theta) = h(t)\,c(\theta)\,e^{w(\theta)t}$ when $w(\theta)$is a nondecreasing function.)

Now, we state the following theorem:

Karlin-Rubin Theorem With H_0: $\theta \leq \theta_0$ and H_1: $\theta > \theta_0$. Suppose that T is a sufficient statistic of θ and its distribution has a nondecreasing MLR. Then, for any t_0, rejecting H_0 iff $T > t_0$ is a UMP level α test, where $\alpha = P_{\theta_0}(T > t_0)$.

In laymen's terms: when the conditions are met, a hypothesis about a parameter above or below the threshold can be translated into a sufficient statistic above or below a threshold. So this is actually very intuitive, maybe even trivial, stuff.

When UMP Cannot Be Found

In case of two-sided hypothesis, the above theorem is not applicable. For example; drawing from n (θ, σ^2), σ^2 known. Say $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$. We look at two alternative values of θ, $\theta_1 < \theta_0 < \theta_2$. We are able to show that different tests are optimal for θ_1 and θ_2, while having the same level. Details:

- Test 1 rejects H_0 when $\bar{X} < \theta_0 - \sigma z_\alpha/\sqrt{n} \iff Z < -z_\alpha - \frac{\sqrt{n}(\theta-\theta_0)}{\sigma}$
- Test 2 rejects H_0 when $\bar{X} > \theta_0 + \sigma z_\alpha/\sqrt{n} \iff Z > z_\alpha + \frac{\sqrt{n}(\theta_0-\theta)}{\sigma}$

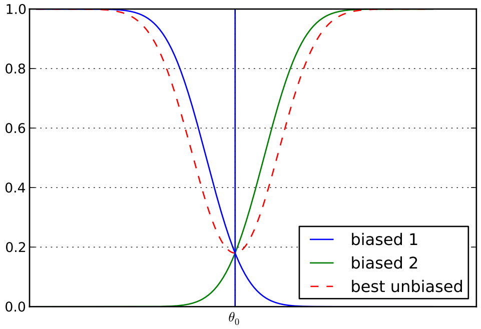
Note they are pathological devices, only controlling for Type I error for the sake of the proof; both tests have spectacular Type II errors. Anyways, as $\bar{X} \sim$ n $\left(\theta, \frac{\sigma^2}{n}\right)$, $\left(\bar{X} - \theta\right) \cdot \frac{\sqrt{n}}{\sigma}$ is going to be a standard normal variable. Therefore you can imagine both tests having level α. However, we can easily see the power for Test 1 would be higher for $\theta_1 < \theta_0$ than Test 2, and vice versa for $\theta_2 > \theta_0$. So there are no UMPs for level α.

Unbiased UMPs

Note the above tests are pathetically biased. What we should do is rejecting H_0 when

$$\left|\bar{X} - \theta_0\right| > \sigma z_{\alpha/2}/\sqrt{n}$$

Obviously. The below figure shows the power function of the three tests:



Note that the best unbiased estimator has a lower probabilty of rejecting H_0 for some range of θ_0, but it is clear it is better than both tests in general.

8.3.3 Size of UIT and IUT

The size/level analysis of UIT and IUT are different. UITs are universally less powerful than LRTs, but they might be easier to reason about.

8.3.4 p-Values

The definition of p-value is very unintuitive, actually. In the definition, a p-value $p(X)$ is a test statistic. Small values of $p(X)$ give evidence that H_1 is true. A p-value is *valid* if, for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$,

$$P_\theta\left(p(X) \leq \alpha\right) \leq \alpha$$

So WTF does this mean? $p(X) \leq \alpha$ can be interpreted as the condition "p-value having a value of α, or something more extreme". The probability of this happening given a θ in the null hypothesis should roughly be α. So this sort of resonates with the colloquial definition of p-value.

Also note that a test that rejects H_0 when $p(X) \leq \alpha$ is a level α test from the above definition.

Defining Valid p-Values

I guess this is the more intuitive definition: let $W(X)$ be a test statistic such that large values of W give evidence that H_1 is true. Then let

$$p(x) = \sup_{\theta \in \Theta_0} P_\theta (W(X) \geq W(x))$$

so this is the colloquial definition. We can prove that $p(x)$ is a valid p-value according to the above definition.

Another method for defining p-values is discussed as well that depends on sufficient statistics. If $S(X)$ is a sufficient statistic of θ when the null hypothesis is true, we have:

$$p(x) = P(W(X) \geq W(x) | S = S(x))$$

is a valid p-value. This makes sense as well; if we have a sufficient statistic we need no stinking inference about θ....

Calculating p-Values

Calculating the supremum probability isn't always easy; but it can be derived from the properties of the distribution. Our usual normal tests are good examples. In these cases, the test statistic follows Student's t distribution, enabling us to nail the supremum.

8.3.5 Loss Function Optimality

Since there are just two actions in a hypothesis test (we either accept H_0 or H_1), a loss function is defined simply by:

$$L(\theta, a_0) = \begin{cases} 0 & \theta \in \Theta_0 \\ C_{II} & \theta \in \Theta_0^C \end{cases} \qquad L(\theta, a_1) = \begin{cases} C_I & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^C \end{cases}$$

where a_i is the action of accepting H_i. C_I and C_{II} are the costs of type I and type II errors, respectively, and they can be a function of θ as well.

9 Interval Estimation

9.1 Introduction

Interval estimation techniques give us confidence intervals of parameters; for example, we can assert $\mu \in [L(x), U(x)]$ where $L(x)$ and $U(x)$ are functions of the random sample that predicts the parameter. What do we gain from going to interval estimations from point estimations? The probability of a point estimation being correct is 0, if the parameter is continuous. However, once we have an interval estimation, we can have nonzero probability of our prediction being correct.

Here are some definitions:

- A coverage probability of $[L(X), U(X)]$ is a function of θ, that represents the probability that the interval covers the true parameter.

$$P_\theta \left(\theta \in \left[L\left(X\right), U\left(X\right)\right]\right)$$

Note that the interval estimates are functions of X, thus are random variables themselves.

- A confidence coefficient is the infimum of coverage probability across all possible

$$\inf_\theta P\left(\theta \in \left[L\left(X\right), U\left(X\right)\right]\right)$$

I want to emphasize the consequences of this definition: don't think interval estimation in terms of conditional probability! A coverage probability is **not** related to the conditional probability. Personally conditional probability of θ given $\bar{X}$ seemed like a very obvious way for interval estimation, but it requires a prior distribution of θ and is strictly not a frequentist approach. The coverage probability is more like a likelihood function. It's just

$$\textit{ConveragePr ob}\left(\theta\right) = \int_C f\left(x, \theta\right) dx$$

where f is the pdf and $C = \left\{x | L\left(x\right) \leq \theta \leq U\left(x\right)\right\}$. This also changes how we interpret the results of an estimation process: the claim that "95% confidence interval of θ is $[1, 10]$" does **not** mean $P\left(\theta \in [1, 10]\right) = 0.95$, but $P\left(x|\theta\right) \geq 0.95$ for all $\theta \in [1, 10]$.

9.1.1 Coverage Probability Example: Uniform Scale Distribution

In many cases, the coverage probability is constant across different parameters. However, it may not be so in certain cases. The following example demonstrates this. Say we draw $X_i \sim \text{uniform}\left(0, \theta\right)$. The sufficient statistic for θ is $Y = \max X_i$. Now, what are our estimates for θ? We examine two candidates:

- Scale intervals:$[aY, bY]$ where $1 \leq a < b$
- Location intervals: $[Y + c, Y + d]$ where $0 \leq c < d$

Now let's send that θ to infinitely. Location interval will have a coverage probability of 0, since the size of the interval stays constant regardless of θ. However, the scale interval manages to have a positive coverage probability because the interval's size roughly *grows* with θ. (These are more intuitive descriptions; formally they are proven by integrating pdfs of Y/θ. Also see the later section on pivotal quantities.)

9.2 Methods of Finding Interval Estimators

9.2.1 Equivalence of Hypothesis Test and Interval Estimations: Inverting a Test Statistic

Hypothesis tests and interval estimations are obviously, very closely related. They both take a random sample to make an inference about the population parameter. To draw a poor analogy: both processes throw darts to a board. Hypothesis testing tell you the target position to shoot for, then you throw the dart, and we will know if we landed within a certain range from the target. Interval estimation lets you throw the dart first; and gives us target positions you might have shot for.

Here's a general recipe for converting a hypothesis test into an interval estimation; just go over all the possible values of θ, use a test of size α for $H_0 : \theta = \theta_0$ where θ_0 is a possible value. If this test passes, θ_0 falls within the confidence interval. So:

$$C\left(T\left(x\right)\right) = \left\{\theta : P\left(T\left(x\right)|\theta\right) \geq \alpha\right\}$$

Of course, in real life, this is not done by enumerating all possible values but analytically.

Normal Test Example

Let $X_i \sim \text{n}\left(\mu, \sigma^2\right)$. Consider testing $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$. Hypothesis test draws a region on the dart, called the acceptance region $A\left(\mu_0\right)$, upon which we will accept H_0. Let's use a test with acceptance region $\left\{\mathbf{x} : |\bar{x} - \mu_0| \leq z_{\alpha/2}\sigma/\sqrt{n}\right\}$. This test has a size of α.

Now, let's "throw the dart" and get a concrete $\bar{x}$. Now, we "wiggle" μ_0 so its acceptance region will still contain $\bar{x}$. It's easy to see the acceptance region will contain $\bar{x}$ when

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Yay, this is a range of parameters. Now what is the confidence coefficient? From the definition of the test, we know

$$P\left(x \in R\left(\mu_0\right)|\mu = \mu_0\right) = \alpha \iff P\left(x \in A\left(\mu_0\right)|\mu = \mu_0\right) = 1 - \alpha$$

So it's $1 - \alpha$!

Formal Definition (Theorem 9.2.2)

For each $\theta_0 \in \Theta$, let $A\left(\theta_0\right)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. For each $x \in \mathcal{X}$, define a set $C\left(x\right)$ in the parameter space by

$$C\left(x\right) = \left\{\theta_0 : x \in A\left(\theta_0\right)\right\}$$

Then the random set $C\left(x\right)$ is a $1 - \alpha$ confidence set. The converse holds true as well; if C is a $1 - \alpha$ confidence set then

$$A\left(\theta_0\right) = \left\{x : \theta_0 \in C\left(x\right)\right\}$$

is a valid acceptance region for a level α test.

Shapes of Confidence Interval

Also, note the above definition do not include the form of H_1. In general, one-based H_1 produces one-sided $C\left(x\right)$ and two-sided H_1 produces two-sided $C\left(x\right)$. The biased property of the inverted test also carries over to the confidence set, as can be expected.

What do we do if we want $C\left(x\right) = \left(-\infty, U\left(x\right)\right]$ i.e. only upper-bounded? If a test uses $H_1 : \mu < \mu_0$, its acceptance region will look like $\left\{x : \bar{x} < \mu_0 - t\right\}$ where t is a function of various things (sample size, nuisance parameters, etc..). i.e. the acceptance region is left-bounded if looked from μ_0. Now, once we have $\bar{x}$, we can move μ_0 infinitely to the left, but cannot move μ_0 too far to the right. So the confidence interval is one-sided only with a upper bound.

9.2.2 Using Pivotal Quantities

A pivotal quantity is a random variable whose distribution does not depend on the parameter. For example, if $X_i \sim \mathrm{n}\left(\mu, \sigma^2\right)$ with σ^2 known, $\left(\bar{\bar{\bar{X}}} - \mu\right) / \left(\sigma/\sqrt{n}\right)$ is normally distributed and is a pivotal quantity. Then, we can trivially construct a confidence interval for μ. First note that

$$P\left(-a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq a\right) = P\left(-a \leq Z \leq a\right) = \alpha$$

Where a is chosen to make the last equality hold. So what μ will make this inequality hold? It's easy to see the following will do it.

$$\left\{\mu : \bar{x} - a\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + a\frac{\sigma}{\sqrt{n}}\right\} = \left\{\mu : \bar{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$

Note that we choose $a = z_{1-\alpha/2}$ which splits the probability equally here. It's a matter of choice, but for symmetric distributions such as normal, it seems natural. For asymmetric distributions such as chi-squared, a different approach will be better (we will revisit this).

In General

If $Q\left(x, \theta\right)$ is a pivotal quantity, we can a and b so that $Q\left(x, \theta\right)$ will fall in that interval with a probability of α (we can plug in concrete numbers, because we know Q's distribution). Then, we have:

$$C\left(x\right) = \left\{\theta_0 : a \leq Q\left(x, \theta_0\right) \leq b\right\}$$

Now, we unravel Q's definition and leave only the desired parameter in the middle, moving everything to the left and right sides. Then we have a confidence interval.

9.2.3 Pivoting CDFs Using Probability Integral Transformation

The previous strategies can sometimes result in non-interval confidence sets, as shown in example 9.2.11. This strategy, when applicable, will always result in an interval. Say $T\left(x\right)$ is a sufficient statistic for θ with cdf $F_T(t|\theta)$. Recall, from probability integral transform, that $F_T(T|\theta) \sim$ uniform $(0,1)$. So this is a pivotal quantity!

For the confidence set to be an interval, we need to have $F_T(t|\theta)$ to be monotone in θ. If it is monotonically increasing, we set $\theta_L(t)$ and $\theta_U(t)$ to be solution of the following system:

$$F_T(t|\theta_L(t)) = \alpha_1 \qquad F_T(t|\theta_U(t)) = 1 - \alpha_2$$

where $\alpha_1 + \alpha_2 = \alpha$. (We usually set $\alpha_1 = \alpha_2 = \alpha/2$ unless additional information. We will discuss about this in section 9.3) Note the above equations do not have to be solved analytically – we can solve them numerically as well.

9.2.4 Bayesian Inference

As noted in the beginning of the chapter, the frequentist approach does not allow to say the parameter belongs to the confidence set with a probability of α: parameter is a fixed value and it will belong to the set with a probability of 1 or 0. The Bayesian setup precisely allows us to do that.

The Bayesian equivalent of a confidence set is called a credible set, to avoid the confusion between the two approaches. The equivalent of a coverage probability is called a credible probability.

In many cases, each approach looks poor if examined using the criteria from a different approach. The textbook shows some examples where the credible set has a limiting converage probability of 0, and the confidence set having a limiting credibility of 0.

9.3 Methods of Evaluating Interval Estimators

We can have multiple ways of coming up with interval estimation, with the same confidence coefficient. Some is bound to be better than the others. But how do we evaluate them?

9.3.1 Size and Coverage Probability

When two interval estimations have the same confidence coefficient, it is obvious to prefer a "smaller" interval. When we estimate the population mean of a normal distribution, we can pick any $\alpha_1, \alpha_2 \geq 0$ where $\alpha = \alpha_1 + \alpha_2$ such that

$$P(z_{\alpha_1} \leq Z \leq z_{\alpha_2}) = 1 - \alpha$$

Obviously there are multiple choices for those. Intuition tells us that splitting probabilities evenly for the left and right hand side of the sample mean is the way to go, and it indeed is. It is justified by the following theorem:

Theorem 9.3.2 Let $f(x)$ be a unimodal pdf. We consider a class of intervals $[a,b]$ where $\int_a^b f(x)\,dx = 1 - \alpha$. If $f(a_0) = f(b_0) > 0$, and $x* \in [a_0, b_0]$ where $x*$ is the mode of f, then

$[a_0, b_0]$ is the shortest interval which satisfies the $1 - \alpha$ probability constraint.

The textbook discusses a caveat for this (See example 9.3.4)

9.3.2 Test-Related Optimality

Optimality criteria for tests carry over to their related interval estimations. For example, we have some guarantee for interval estimations which are inversions of UMP tests. However, note that UMP controls type-II errors; therefore the guarantee we have for the interval estimation also looks different. UMP-inverted estimations give us optimal *probability of false coverage*. Colloquially, it is a function of θ and θ' and measures the probability of θ' being covered when θ is the truth.

Note the probability of false coverage has to take different forms according to the form of the original test. Say the the interval estimation is one-sided only with upper bound. We only define the probability of false coverage for $\theta < \theta'$, obviously – θ' less than or equal to θ are correct.

Intuition Behind False Coverage Probability We do offer some intuition about the linkage between the false coverage probability and the size of the interval; if an interval contains less false parameters, it's more likely to be short; there are some loose links discussed in the book (see theorem 9.3.9) which says the expected length of the interval for a given parameter value is equivalent to an integral over false positive probability.

Uniformly Most Accurate (UMA) Intervals A conversion of a UMP test yields a UMA confidence set, which has the smallest probability of false coverage.

Also note that UMP tests are mostly one-sided – Karlin-Rubin works only for one sided intervals. So most UMA intervals are one sided.

Inverting Unbiased Tests The biasedness of tests carry over to interval estimation. In case UMP does not exist, we can invert the unbiased test to get an unbiased interval.

9.3.3 Bayesian Optimality

When we have a posterior distribution, we can order all intervals of probability of $1 - \alpha$ by their size. A corollary to Theorem 9.3.2 mentioned above gives us that when the posterior distribution is unimodal,

$$\{\theta : \pi(\theta|x) \geq k\}$$

is the shortest credible interval for a credible probability of $1 - \alpha = \int_{\pi(\theta|x) \geq k} \pi(\theta|x)\, dx$. Such a region is called the highest posterior density (HPD) region.

9.3.4 Loss Function Optimality

So far we have set the minimum coverage probability we want, and then find the best tests among them. We can do something in the middle by using a generalized loss function such as

$$L(\theta, C) = b \cdot \textit{Length}(C) - I_C(\theta)$$

where $I_C(\theta) = 1$ if $\theta \in C$, 0 otherwise. By varying b, we can very the relative importance of the length and coverage probability.

However, the use of decision theory in interval estimation problems is not widespread; it is hard to solve in many cases, and this can sometimes lead to unexpected types of sets.

10 Asymptotic Evaluations

This chapter looks at asymptotic behaviors of several topics (point estimation, hypothesis testing, and interval estimation). That is, we send sample size to infinity and see what happens. The detailed treatment of the topic seems to be the most theoretical part of the book; I just skimmed over the chapter and will try to summarize only the "big" ideas here.

10.1 Point Estimation

10.1.1 Criteria

A point estimator's asymptotic behavior is characterized by two properties;

- *Consistency* means the estimator converges to the "correct" value as the sample size becomes infinite.

  $$\lim_{n\to\infty} P_\theta(|W_n - \theta| < \epsilon) = 1$$

- *Efficiency* looks at the variance of the estimator as the sample size becomes infinite. If the variance reaches the lower bound defined by the Cramer-Rao, the estimator is called asymptotically efficient.

  Concretely, an asymptotically efficient estimator W_n for a parameter $\tau(\theta)$ if

  $$\sqrt{n}\,[W_n - \tau(\theta)] \to \text{n}\,[0, v(\theta)]$$

  and $v(\theta)$ achieves the Cramer-Rao Lower Bound.

It is notable that MLEs are in general best estimators: they are consistent estimators, and asymptotically efficient at the same time. (Some "regularity" conditions are required, but it is explicitly said to hold in most common situations and you don't really want to care about it.)

10.1.2 Comparing Consistent Estimators

MLE is everyone's favorite estimator. However, other estimators may have other desirable properties (robustness, ease of calculation, etc) so we need to be able to see what we are giving up in terms of efficiency. Comparing different asymptotically consistent estimators is done by looking at their variances, through the idea of *asymptotic relative efficiency* (ARE). If two estimators W_n and V_n satisfy

$$\begin{aligned} \sqrt{n}\left[W_n - \tau\left(\theta\right)\right] &\rightarrow \text{n}\left[0, \sigma_W^2\right] \\ \sqrt{n}\left[V_n - \tau\left(\theta\right)\right] &\rightarrow \text{n}\left[0, \sigma_V^2\right] \end{aligned}$$

Then, the ARE is defined by

$$\text{ARE}\left(V_n, W_n\right) = \frac{\sigma_W^2}{\sigma_V^2}$$

ARE is going to be a function of the parameters – so we will be able to see where it peaks, where is it larger/smaller than 1, etc.

10.1.3 Asymptotic Behavior of Bootstrapping

Bootstraps are good ways to estimate variances of arbitrary estimators. For any estimator $\hat{\theta}\left(x\right) = \hat{\theta}$,

$$\text{Var}^*\left(\hat{\theta}\right) = \frac{1}{n^n - 1}\sum_{i=1}^{n^n}\left(\hat{\theta}_i^* - \bar{\hat{\theta}}^*\right)^2$$

Of course, we cannot sample all n^n possible samples. So we can always do a partial bootstrap by taking B random resamples. The book shows some examples where bootstrap yields a better variance estimate than the Delta method (which exploits that when estimates are asymptotically efficient, they will reach Cramer-Rao bound – so naturally it will underestimate.).

Parametric and Nonparametric Bootstrap

The usual type of bootstrapping, which draws data (with replacement) to generate random samples is called a nonparametric bootstrap. On the contrary, *parametric bootstrapping* assumes a distribution. Say $X_i \sim f\left(x|\theta\right)$. We take MLE estimate $\hat{\theta}$, and generate random samples from there.

$$X_1^*, X_2^*, \cdots, X_n^* \sim f\left(x|\hat{\theta}\right)$$

It is same as the usual nonparametric bootstrap from there.

Consistency and Efficiency

The textbook does not cover a lot of material about the evaluation of bootstrapping. In general, it is an effective and reasonable way.

10.2 Robustness

1. It should have a reasonably good efficiency at the assumed model.
2. It should be robust in the sense that small deviations from the model assumptions should impair the performance only slightly.
3. Somewhat larger deviations from the model should not cause a catastrophe.

10.2.1 Robustness of Mean and Median

The mean reaches the Cramer-Rao bound, so it is an efficient estimator. Also, if there is a small variation from the normal distribution assumption, it will fare pretty well. We can try to see this by a δ-contamination model. There, the distribution is the assumed model with a probability of $1-\delta$, some other distribution with δ. Note when the two distributions are similar, mean's variance is still small. However, if the other distribution is Cauchy, for example, the variance will go to infinity.

That brings us to the notion of breakdown value. A breakdown value b is the maximal portion of the sample which can go to infinity before the statistic goes infinity, or "meaningless". Of course, the mean's breakdown value is 0, where the median has 0.5.

How can we compare the mean and the median? We can prove median is asymptotically normal, and use ARE to compare those in different types of distributions. Some examples in the book show it will fare better than the mean in double exponential distribution, but not in normal or logistic. So it fares better with thicker tails, as we expect.

10.2.2 M-estimators

M-estimators is a generalized form of estimator. Most estimators minimize some type of criteria - for example, squared error gives the mean, absolute error gives the median, and the negative log likelihood will give you MLE. M-estimators are a class of estimators which minimize

$$\sum_{i=1}^{n} \rho\left(x_i - a\right)$$

Huber's Loss

Huber's loss is a Frankenstein-style loss function created by patching squared and absolute loss together.

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le k \\ k|x| - \frac{1}{2}k^2 & \text{otherwise} \end{cases}$$

Works like quadratic around zero, and linear away. It's differentiable, and continuous. k is a tunable parameter. Increasing k would be like decreasing the robustness to outliers. The minimizer to this loss function is called the Huber estimator and it is asymptotically normal with mean θ. If you do an ARE comparison with mean and median;

- Huber is close to mean in normal, and better than mean at logistic or double exponential.
- Huber is worse than median in double exponential, much better than it in normal or logistic.

10.3 Hypothesis Testing

10.3.1 Asymptotic Distribution of LRT

When $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$. Suppose $X_i \sim f(x|\theta)$. Then under H_0, as $n \to \infty$,

$$-2\log\lambda(X) \to \chi_1^2 \text{ in distribution}$$

Regardless of the original distribution! Kicks ass. A more general version, which do not specify H_0 and H_1 explicitly, states that the quantity $-2\log\lambda(X)$ will still converge to chi-squared distribution with its df equal to the difference between the number of free parameters in Θ_0 and the number of free parameters in Θ_1.

How do we define the number of free parameters? Most often, Θ can be represented as a subset of $\mathbf{R}^q$ and Θ_0 can be represented as a subset of $\mathbf{R}^\mathbf{p}$. Then $q - p = v$ is the df for the test statistic.

10.3.2 Wald's Test and Score Test

Two more types of tests are discussed.

Wald Test

When we have an estimator W_n for a parameter θ, which will asymtotically normal, we can use this as a basis for testing $\theta = \theta_0$. In general, a Wald test is a test based on a statistic of the form

$$Z_n = \frac{W_n - \theta_0}{S_n}$$

S_n is standard error for W_n. (Last time I've seen this was in logistic regression; they did Wald Test on regression coefficients to derive p-values of coefficient being nonzero.)

Score Test

10.4 Interval Estimation

11 Analysis of Variance and Regression

11.1 One-way ANOVA

ANOVA is a method of comparing means of several populations, often assumed to be normally distributed. Normally, the data are assumed to be follow the model

$$Y_{ij} = \theta_i + \epsilon_{ij}$$

where θ_i are unknown means and ϵ_{ij} are error random variables. The classic oneway ANOVA assumptions are as follows:

1. $\mathrm{E}\epsilon_{ij} = 0$, $\mathrm{Var}\epsilon_{ij} = \sigma_i^2 < \infty$, and all errors are uncorrelated.

2. ϵ_{ij} are independent, normally distributed

3. $\sigma_i^2 = \sigma^2$ for all i (also known as *homoscedascity*)

11.1.1 Different ANOVA Hypothesis

The classic one-way ANOVA null hypothesis states

$$H_0 : \theta_1 = \theta_2 = \theta_3 = \cdots = \theta_k$$

which is kind of uninformative, and doesn't give us much information. The text starts with a more useful type of hypothesis, using *contrasts*. A contrast is a linear combination of variables /parameters where the weights sum up to 0. We now run the test with different hypothesis, the null being

$$H_0 : \sum_{i=1}^{k} a_i \theta_i = 0$$

Now, by choosing the weights carefully, we can ask other types of interesting questions. $a = (1, -1, 0, 0, \cdots, 0)$ will ask if $\theta_1 = \theta_2$. $a = \left(1, -\frac{1}{2}, -\frac{1}{2}, 0, 0, \cdots, 0\right)$ will ask if $\theta_1 = (\theta_2 + \theta_3)/2$, etc.

11.1.2 Inference Regarding Linear Combination of Means

The means of each sample $\bar{Y}_i$ are normal:

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \sim \mathrm{n}\left(\theta_i, \sigma^2/n_i\right)$$

The linear combination of normal variables are once again, normal, with the following parameters:

$$\sum_{i=1}^{k} a_i \bar{Y}_i \sim \mathrm{n}\left(\sum_{i=1}^{k} a_i \theta_i, \sigma^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}\right)$$

Since we don't know the variance, we can replace this as the sample variance. S_i^2 is the regular sample variance from the ith sample. Then, the pooled estimator is given by

$$S_p^2 = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1) S_i^2 = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$$

The estimator S_p^2 has an interesting interpretation: it is the mean square within treatment groups. This makes sense, since σ^2 only affects the variance within a group in our model. Then, replacing the variance with S_p^2, we get the following which has usual Student's t-distribution with $N-k$ degrees of freedom.

$$\frac{\sum_{i=1}^{k} a_i \bar{Y}_i - \sum_{i=1}^{k} a_i \theta_i}{\sqrt{S_p^2 \sum_{i=1}^{k} a_i^2 / n_i}} \sim t_{N-k}$$

Now, we can do the usual t-test. In cases where we are only checking the equivalence of two means, this is equivalent to the two-sample t test is that here information from other groups are factored in estimating S_p^2.

11.1.3 The ANOVA F Test

How do we test the classic hypothesis? We can think of it as an intersection of multiple hypothesis:

$$\theta \in \{\theta : \theta_1 = \theta_2 = \cdots = \theta_k\} \iff \theta \in \Theta_a \text{ for all contrast a} \iff \theta \in \cap_{a \in \mathcal{A}} \Theta_a$$

We can reject this intersection if the test fails for any a. We can test an individual hypothesis with H_{0a} by the following statistic:

$$T_a = \left| \frac{\sum_{i=1}^{k} a_i \bar{Y}_i - \sum_{i=1}^{k} a_i \theta_i}{\sqrt{S_p^2 \sum_{i=1}^{k} a_i^2 / n}} \right|$$

where we will reject it if $T_a > k$ for some k. Now, all tests will pass if and only if

$$\sup_{a \in \mathcal{A}} T_a \leq k$$

Where $\mathcal{A} = \{a : \sum a = 0\}$. How can we find this supremum? Lemma 11.2.7 gives us the exact form of a and the value of the supremum. However, the important thing is that:

$$\sup_{a\in\mathcal{A}} T_a^2 = \frac{\sum_{i=1}^{k} n_i \left(\left(\bar{Y}_i - \bar{\bar{Y}}\right) - \left(\theta_i - \bar{\theta}\right)\right)^2}{S_p^2} \sim (k-1)\, F_{k-1,N-k}$$

the statistic of which is called the F-statistic, and it gives us the F-test. Now, we can reject H_0 when

$$\frac{\sum_{i=1}^{k} n_i \left(\bar{Y}_i - \bar{\bar{Y}}\right)^2 / (k-1)}{S_p^2} > F_{k-1,N-k,\alpha}$$

What is the rationale behind looking at F statistic? The denominator is the estimated variance within groups. The numerator is the mean square between treatment groups, weighted by the size of the group. $\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2$ is the squared error between the group mean and the grand mean. n_i weights them by the size of the group. Dividing by $k-1$ is getting the average error per group. Now, the ratio between these two quantities should be higher when inter-group variance is high relative to the intra-group variance.

11.1.4 Simultaneous Estimation of Contrasts

A couple of strategies for making inferences about multiple equalities are discussed; the Bonferroni procedure and the Scheffe's procedure. The Bonferroni procedure allows you to make inferences about m pairs of means being equal. You have to set m in advance, and adjust the level of the test so the intersection tests will be of desired power.

Scheffe's procedure is more notable, which allows you to construct confidence intervals on any arbitrary contrasts after the procedure is done. It is noted as a legitimate use of data snooping. However, at the cost of power of inference, the intervals are usually wider. It goes like:

If $M = \sqrt{(k-1)\, F_{k-1,N-k,\alpha}}$, then the probability is $1-\alpha$ that

$$\sum_{i=1}^{k} a_i \bar{Y}_i - M\sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}} \le \sum_{i=1}^{k} a_i \theta_i \le \sum_{i=1}^{k} a_i \bar{Y}_i + M\sqrt{S_p^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}$$

11.1.5 Partitioning Sum of Squares

ANOVA provides a useful way of thinking about the way in which different treatments affect a measured variable. We can allocate variation of the measured variable to different sources, because:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}}\right)^2 = \sum_{i=1}^{k} n_i \left(\bar{y}_{i\cdot} - \bar{\bar{y}}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\cdot}\right)^2$$

which can easily be proved because $\left(y_{ij} - \bar{\bar{y}}\right)^2 = \left(\left(y_{ij} - \bar{y}_i\right) + \left(\bar{y}_i - \bar{\bar{y}}\right)\right)^2$ and when you evaluate the square, cross terms are zero. The sums of squares are also chi-square distributed, after scaling.

11.2 Simple Linear Regression

Simple linear regression is discussed in three different contexts - as a minimizer to the least-squares without any statistical assumptions, as a best linear unbiased estimators under some variance assumptions, as an inference mechanism under distribution assumptions. Not surprisingly, we will be able to draw more powerful conclusions when we assume more.

11.2.1 General Model

In all three different contexts, the actual line stays the same. Say (x_i, y_i) are pairs of examples, where x_i are the predictor variables, y_i being response variables. Then $\bar{x}$ and $\bar{y}$ are means of x_i and y_i, respectively, and

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ S_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ S_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

And then we will fit the following line $y = bx + a$ with:

$$\begin{aligned} b &= \frac{S_{xy}}{S_{xx}} \\ a &= \bar{y} - b\bar{x} \end{aligned}$$

Note the slope can be interpreted as Cov (X, Y) /Var$X$.

11.2.2 Least Square Solution

Least square provides a way to “fit” a line to the given data. No statistical inferences can be drawn here. Let’s say we want to minimize the residual sum of squares

$$\text{RSS} = \sum_{i=1}^{n} (y_i - (c + dx_i))^2$$

Now c can be determined easily - rewrite the summand as

$$(y_i - (c + dx_i))^2 = ((y - dx_i) - c)^2$$

and the minimizer of this is just the average of $y - dx_i$ which is $\bar{y} - d\bar{x}$. d can be determined from differentiating the quadratic formula and setting it to 0.

Also note that changing the direction of the regression (using y to predict x) will give you a different regression line: this is obvious since b becomes S_{yy}/S_{xy}.

11.2.3 Best Linear Unbiased Estimators: BLUE

Let's add some contexts: we now think of the values y_i as observed values of uncorrelated random variables Y_i. x_i are known, fixed values chosen by the experimenter. We assume the following model:

$$\mathrm{E}Y_i = \alpha + \beta x_i$$

where Var$Y_i = \sigma^2$, which is a *common* variance across variables. Or equivalently, set $\mathrm{E}\epsilon_i = 0$, Var$\epsilon_i = \sigma^2$ and have

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

Now, let us estimate α and β as a linear combination of Y_is

$$\sum_{i=1}^{n} d_i Y_i$$

Furthermore, we only look at unbiased estimators. With an unbiased estimator of slope β must satisfy

$$\mathrm{E}\sum_{i=1}^{n} d_i Y_i = \beta$$

We can transform LHS as

$$\mathrm{E}\sum_{i=1}^{n} d_i Y_i = \sum_{i=1}^{n} d_i \mathrm{E}Y_i = \sum_{i=1}^{n} d_i \left(\alpha + \beta x_i\right) = \alpha\left(\sum_{i=1}^{n} d_i\right) + \beta\left(\sum_{i=1}^{n} d_i x_i\right)$$

For this to be β, we need the following conditions to hold true:

$$\sum_{i=1}^{n} d_i = 0 \text{ and } \sum_{i=1}^{n} d_i x_i = 1$$

So the minimum variance estimator which satisfies the above conditions is called the *best linear unbiased estimator* (BLUE). The d_is which satisfy this could be find using a similar strategy for maximizing T_a in section 11.1.3 above. After the dust settles, we have:

$$d_i = \frac{(x_i - \bar{x})}{S_{xx}}$$

which seems to have an interesting interpretation. Higher S_{xx} make the coefficients smaller, and x_i deviating more from $\bar{x}$ makes coefficients larger.

What is the variance of β now?

$$\mathrm{Var}b = \sigma^2 \sum_{i=1}^{n} d_i^2 = \frac{\sigma^2}{S_{xx}}$$

11.2.4 Normal Assumptions

Now, we can assume normality for the variables which let us make further claims regarding the estimators. The text discusses two ways of doing this, which are practically equivalent. The more common one is the conditional normal model, which states

$$Y_i \sim \text{n}\left(\alpha + \beta x_i, \sigma^2\right)$$

which is a special case of the model discussed above. Even less general is the bivarite normal model, which assumes the pair (X_i, Y_i) follows a bivariate normal distribution. However, in general we don't care about the distribution of X_i, but only the conditional distribution of Y_i. So bivariate normal assumptions are not used often.

Also, note both models satisfy the assumptions we have made in the above section.

Maximum Likelihood

Under this distribution assumption, we can try to find the MLE of β. We expect to find the same formula - and we actually do. The log likelihood function is maximized at the same choice of β and α.

What about the MLE of σ^2? It is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2$$

which is the variance of the error (RSS) - makes sense because RSS are effectively ϵ_i with Var$\epsilon_i = \sigma^2$! However, note that $\hat{\sigma}^2$ is not an unbiased estimator of σ^2.

Distributions of Estimators Under Normality Assumption

The sampling distributions of the maximum likelihood estimates $\hat{\alpha}$, $\hat{\beta}$, and S^2 are

$$\hat{\alpha} \sim \text{n}\left(\alpha, \frac{\sigma^2}{nS_{xx}}\sum_{i=1}^{n} x_i^2\right) \quad \hat{\beta} \sim \text{n}\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

with

$$\text{Cov}\left(\hat{\alpha}, \hat{\beta}\right) = \frac{-\sigma^2 \bar{x}}{S_{xx}}$$

Furthermore, $\left(\hat{\alpha}, \hat{\beta}\right)$ and S^2 are independent and

$$\frac{(n-2)\,S^2}{\sigma^2} \sim \chi^2_{n-2}$$

When σ^2 is unknown, we can still make inferences about them using S^2 since we get Student's t-distribution using:

$$\frac{\hat{\alpha} - \alpha}{S\sqrt{\left(\sum_{i=1}^{n} x_i^2\right) / \left(nS_{xx}\right)}} \sim t_{n-2}$$

and

$$\frac{\hat{\beta} - \beta}{S/\sqrt{S_{xx}}} \sim t_{n-2}$$

Significance of the Slope

The t-test for significance of β will reject $H_0 : \beta = 0$ when

$$\left|\frac{\hat{\beta} - 0}{S/\sqrt{S_{xx}}}\right| > t_{n-2,\alpha/2}$$

This is equivalent to the following, since t-distribution squared is distributed following the F-distribution.

$$\frac{\hat{\beta}^2}{S^2/S_{xx}} > F_{1,n-2,\alpha}$$

The LHS quantity, the F-statistic, can be interpreted as follows:

$$\frac{\hat{\beta}^2}{S^2/S_{xx}} = \frac{S_{xy}^2/S_{xx}}{\text{RSS}/\left(n-2\right)} = \frac{\text{Regression sum of squares}}{\text{Residual sum of squares/df}}$$

which is nicely summarized in an ANOVA table.

Partitioning Sum of Squares

As an another similarity to ANOVA, we can express the total sum of squares in the data set by a sum of regression sum of squares and the residual sum of squares:

$$\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}\right)^2 + \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

When we split the sum of squares, we can take the ratio between the regression sum of squares and the total sum of squares as the *coefficient of determination*, called r^2:

$$r^2 = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

The last portion of this inequality is not very obvious... look at Exercise 11.34 for more intuition.

Prediction at Specified x_0

We are able to discuss the distribution of the response variable at a specified position x_0. Call this Y_0. Under our assumptions, $\mathrm{E}\left(Y|x_0\right) = \alpha + \beta x_0$ which is estimated by $\hat{\alpha} + \hat{\beta} x_0$ which is an unbiased estimator. What is the variance of this estimator?

$$\begin{aligned}
\text{Var}\left(\hat{\alpha} + \hat{\beta} x_0\right) &= \text{Var}\hat{\alpha} + \left(\text{Var}\hat{\beta}\right) x_0^2 + 2x_0 \text{Cov}\left(\hat{\alpha}, \hat{\beta}\right) \\
&= \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} x_i^2 + \frac{\sigma^2 x_0^2}{S_{xx}} - \frac{2\sigma^2 x_0 \bar{x}}{S_{xx}} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{S_{xx}}\right)
\end{aligned}$$

Now we have a normal distribution for Y_0. For inference, the following quantity follows a Student's t-distribution:

$$\frac{\hat{\alpha} + \hat{\beta} x_0 - \left(\alpha + \beta x_0\right)}{S\sqrt{\frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{S_{xx}}}} \sim t_{n-2}$$

which can be used to make confidence interval for $\mathrm{E}Y_0 = \alpha + \beta x_0$.

Prediction Interval

The previous estimation and inference was done on the estimator $\hat{\alpha} + \hat{\beta} x_0$, which is our estimate for the mean of Y_0. Now, can we make intervals for Y_0 itself? Obviously, the interval is going to be larger - we should add variance from the distribution of Y_0 as well.

Here's the definition: a $100\left(1-\alpha\right)\%$ prediction interval for an unobserved random variable Y based on the observed data X is a random interval $\left[L\left(X\right), U\left(X\right)\right]$ such that

$$P_\theta\left(L\left(X\right) \le Y \le U\left(X\right)\right) \ge 1 - \alpha$$

for all θ. The variance of Y_0 is given by summing up the variance of the mean estimator and the common variance σ^2.

Simulataneous Estimation and Confidence Bands

We can create "confidence bands" around the fitted line, which gives us confidence intervals for the mean of Y at that x. This is similar to getting confidence bands in ANOVA, and the same two processes apply: Bonferroni and Scheffe. Without further details, we state the Scheffe band:

Under the conditional normal regression model, the probability is at least $1 - \alpha$ that

$$\hat{\alpha} + \hat{\beta} x - M_\alpha S \sqrt{\frac{1}{n} + \frac{\left(x - \overline{x}\right)^2}{S_{xx}}} < \alpha + \beta x < \hat{\alpha} + \hat{\beta} x + M_\alpha S \sqrt{\frac{1}{n} + \frac{\left(x - \overline{x}\right)^2}{S_{xx}}}$$

where $M_\alpha = \sqrt{2F_{2,n-2,\alpha}}$.

12 Regression Models

The last chapter! Yay! Can't believe I made it so far. (Well, yeah, I skipped over a good amount of material...) Anyways, this chapter covers a number of different models for regression.

12.1 Errors in Variables (EIV) Models

In EIV models, contrary to the traditional regression methods, the xs, as well as ys, are realized values of a random variable whose mean we cannot observe: $\mathrm{E}X_i = \xi_i$. The means of the two families of variables are linked by a linear relationship. If $\mathrm{E}Y_i = \eta_i$,

$$\eta_i = \alpha + \beta\xi_i$$

In this model, there is really no distinction between the predictor variable and the response variable.

12.1.1 Functional And Structural Relationship

There are two different types of EIV models. The more obvious one is the linear "functional" relationship model, where ξ_is are fixed, unknown parameters. Adding more parameterization gives us the linear "structural" relationship model, where $\xi_i \sim \text{iid n}\left(\xi, \sigma_\xi^2\right)$. In practice, they share a lot of properties and the functional model is used more often.

12.1.2 Mathematical Solution: Orthogonal Least Squares

OLS regression measures the vertical distance between each point and the line, since we trust xs to be correct. In EIV, there is no reason to do that and we switch to orthogonal regression. Here, the deviation is the distance between the point and the regression line. The line segment spanning this distance is orthogonal to the regression line, thus the name. The formula for this in case of a simple regression is given in the book.

Orthogonal least squares line always lies between the two OLS regression lines - y on x and x on y.

12.1.3 Maximum Likelihood Estimation

The MLE of the functional linear model is discussed. The obvious likelihood function, however, does not have a finite maximum. (Setting derivatives to zero results in a saddle point.) To avoid this problem, we change the model where we do not know the variances of the two errors (one in x and one in y), but their ratio λ.

$$\sigma_\delta^2 = \lambda\sigma_\epsilon^2$$

Note since $\text{Var}X = \sigma_\delta^2$, this includes the regular regression model when $\lambda = 0 \implies \text{Var}X = 0$. The maximization can be done analytically. This MLE, when $\lambda = 1$, will be the result of the orthogonal least squares. When we send $\lambda \to 0$, it will become the regular OLS results. Cool right?

Case for the structural model is discussed, but I'm going to just skip over it.

12.1.4 Confidence Sets

Omitted.

12.2 Logistic Regression And GLM

12.2.1 Generalized Linear Model

A GLM consists of three components: the random component (response variables), the systematic component (a function $h(x)$ of predictor variables, linear *in the parameter*), and the link function $g(\mu)$. Then the model states

$$g(\mathrm{E}Y_i) = h(x)$$

Important points: the response variables are supposed to come from a specified *exponential* family.

12.2.2 Logistic Regression

In logistic regression, $Y_i \sim \text{Bernoulli}(\pi_i)$, g is the logit function. Let us limit h to have the form $\alpha + \beta x_i$ for easier discussion. Then we have

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i$$

Note that $\log(\pi/(1-\pi))$ is the natural parameter of the Bernoulli family, since the pmf can be represented as

$$\pi^y (1-\pi)^{1-y} = (1-\pi)\exp\left\{y \log\left(\frac{\pi}{1-\pi}\right)\right\}$$

when the natural parameter is used as the link function, as in this case, it is called the canonical link. We can rewrite the link equation which gives us better intution about how the probability and the linear combination is related.

$$\pi_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$$

Estimating logistic regression is done by MLE, as we don't have a clear equivalent of least squares. This will be done numerically.

12.3 Robust Regression

Remind the relationship between mean (minimizes L2 mean) and the median (minimizes L1 mean). There is a median-equivalent for least squares; which is called LAD (least absolute deviation) regression. It minimizes

$$\sum_{i=1}^{n} |y_i - (a + bx_i)|$$

This is L1 regression. (Solvable by LP.) As can be expected, it is quite more robust against outliers. However, the asymptotic normality analysis gives as the ARE of least squares and LAD is $4f(0)^2$ (f is the standard normal pdf) which gives us about 64%. So we give up a good bit of efficiency against least squares.

12.3.1 Huber Loss

Analogeous to M-estimator, we can find regression functions that minimizes the Huber loss. The analysis of this is complicated and is omitted from the book as well. However, it hits a good middle ground between the two extreme regression techniques. The book demonstrates this over three datasets; where the errors are generated from normal, logistic, and double exponential distributions. Then, the AREs are calculated between least squares, LAD and M-estimator. The result is very good. Here I replicate the table:

Error	Normal	Logistic	Double Exponential
vs. least squares	0.98	1.03	1.07
vs. LAD	1.39	1.27	1.14

Almost as good as least squares in normal, completely kick arse in other cases. Very impressive!! Also, note LAD is worse off than least squares in everything. What a piece of crap. Anyways, I sort of understand why professor Boyd said Huber loss will improve things greatly!