Point Estimation
(Devore Chapter Six)

MATH-252-01: Probability and Statistics II*

Spring 2019

Contents

Tuesday 15 January 2019

0 Preliminaries

0.1 Administrata

- Syllabus
- Instructor's name (Whelan) rhymes with "wailin'".
- Text: Devore, *Probability and Statistics for Engineering and the Sciences*. The official version is the 9th edition, which is probably the version you used in MATH 251. There are older editions around, and the changes between editions are generally minimal, but make sure you're doing the right version of assigned problems! (We won't be using WebAssign for homework.) The 7th and 8th editions are also on reserve at the library.
- Course website: `http://ccrg.rit.edu/~whelan/MATH-252/`. I intend to post materials there rather than on mycourses.
- Course calendar: *tentative* timetable for course.
- Structure:

- Read relevant sections of textbook before class
- Lectures to reinforce and complement the textbook
- Practice problems (odd numbers; answers in back but more useful if you try them before looking!).
- Problem sets to hand in: practice at writing up your own work neatly & coherently. Problem sets will also contain some numerical exercises intended to be done in minitab. Note: doing the problems is *very* important step in mastering the material.
- Minitab: proprietary statistical software package, sort of like Excel with built-in statistical functionality. Available for free-as-in-beer download on campus (or over VPN) via `https://www.rit.edu/its/services/software-licensing/minitab`. Primary version runs under Windows; there is also "Minitab Express" for Mac (and Windows). Apparently no version exists for Linux, and I haven't been able to get either one to run under an emulator. Minitab is installed in all of the computer labs, including Gosnell 08-1345. It will probably be possible (if less straightforward) to do the numerical exercises in another environment, like Python/SciPy, although someone is likely to expect you to know minitab down the road.
- Quizzes: closed book, closed notes, use *scientific* calculator (not graphing calculator, *not* your phone!)
- Prelim exams (think midterm, but there are two of them) in class at roughly 1/3 and 2/3 of the way through the course: closed book, one handwritten formula sheet, use scientific calculator (*not* your phone!)
- Final exam will be cumulative (but focus more on last third of the course).

- Grading:

  5% Problem Sets & Computer Exercises
  10% Quizzes
  25% First Prelim Exam
  25% Second Prelim Exam
  35% Final Exam

  You'll get a separate grade on the "quality point" scale (e.g., 2.5–3.5 is the B–including B+ and B−–range) for each of these five components; course grade is weighted average.

0.2 Outline

1. Parameter Estimation
   (a) Point Estimation (Chapter Six)
   (b) Interval Estimation (Chapter Seven)
2. Hypothesis Testing
   (a) One-Sample Hypothesis Testing (Chapter Eight)
   (b) Two-Sample Inference (Chapter Nine)
3. Model Fitting
   (a) Regression (Chapter Twelve)
   (b) Goodness of Fit (Chapter Fourteen)
4. Non-Parametric Methods (Chapter Fifteen, time permitting)

Warning: you will generally be expected to recall and apply what you learned in MATH 251. For convenience, these notes include a short review of some of the most relevant parts, for your perusal outside of class.

0.3 Review of Statistical Formalism

0.3.1 Descriptive Statistics

In this course, we will perform a number of manipulations on data sets in order to make probabilistic statements on the underlying source of the data. (E.g., properties of a population from which a sample may be drawn.) The basic building blocks of these calculations are the quantities of descriptive statistics, covered in Chapter One of Devore (see `http://ccrg.rit.edu/~whelan/courses/2013_1sp_1016_345/notes01.pdf` for more details.)

As a quick refresher, consider rainfall totals[1] from a weather station in Phoenix, AZ for the years 2011-2015: 4.92, 5.35, 6.77, 8.74, and 5.08 inches, respectively. We write this as $x_1 = 4.92$, $x_2 = 5.35$, $x_3 = 6.77$, $x_4 = 8.74$, and $x_5 = 5.08$ inches, respectively. Recall some of the basic summary statistics we can construct from these data:

- To get the *sample median* $\widetilde{x}$, we sort the values in order from lowest to highest, and pick the middle one:

  $$4.92,\ 5.08,\ 5.35,\ 6.77,\ 8.74$$

  Thus $\widetilde{x} = 5.35$. Note that at least half the $\{x_i\}$ have $x_i \leq \widetilde{x}$ and at least half have $x_i \geq \widetilde{x}$. The median is also called the 50th percentile, and this can be extended to other choices: 6.77 is the 70th percentile because at least 70% of the values have $x_i \leq 6.77$, and at least 30% have $x_i \geq 6.77$.

[1] `http://alert.fcd.maricopa.gov/alert/Rain/Master/4810.pdf` Note that we've taken a rather small dataset to illustrate what's happening in these calculations by hand. In practice, you'd process any decent-sized dataset with a computer package of some sort.

- The *sample mean* $\overline{x}$ is the average value

  $$\begin{aligned}\overline{x} &= \frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5) = \frac{1}{5}\sum_{i=1}^{5} x_i \\ &= \frac{4.92 + 5.35 + 6.77 + 8.74 + 5.08}{5} = \frac{30.86}{5} = 6.172\end{aligned} \tag{0.1}$$

  Note that it would be more appropriate to quote this value as 6.17, because the individual values are quoted to three *significant figures*, but we don't know that e.g., 4.92 means 4.920000000 and not 4.924 or 4.916. As a general rule, your answers shouldn't carry more significant figures than the experimental data you start with. Your calculator, statistical software program, etc can carry more than that, and it's good to keep some extra digits for internal calculations and not round off intermediate quantities too much.

  In general, if there are n data points in the sample, the sample mean is defined as $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

- The *sample variance* s^2 is defined as

  $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{0.2}$$

  It's sort of an average square deviation from the sample mean (we'll get to the reason it's $n-1$ rather than n in a moment). So to construct it for our rainfall data, we'd do the following:

i	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
1	4.92 in	−1.252 in	1.567504 in^2
2	5.35 in	−0.822 in	0.675684 in^2
3	6.77 in	0.598 in	0.357604 in^2
4	8.74 in	2.568 in	6.594624 in^2
5	5.08 in	−1.092 in	1.192464 in^2

Adding the last column gives $10.38788\,\text{in}^2$, so the sample variance in this case is $s^2 = \frac{10.38788\,\text{in}^2}{4} = 2.59697\,\text{in}^2$. Note the units on this are inches-squared, not inches. If we write this to two significant figures, we get $2.60\,\text{in}^2$.

- The sample standard deviation s is the square root of the sample variance, so here $s = \sqrt{2.59697\,\text{in}^2} \approx 1.61\,\text{in}$.

There is a mathematical trick that notes that (after some algebra)

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i)^2 - n(\overline{x})^2 \tag{0.3}$$

which can be used to calculate the sample variance as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i)^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}x_i\right)^2 \tag{0.4}$$

in the case where you happen to know $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n}(x_i)^2$. This "shortcut" is actually somewhat dangerous with real data, though; if it happens that the typical value of $(x_i - \overline{x})^2$ is a lot smaller than $\overline{x}^2$ itself, you can have a situation where the two terms being subtracted in (0.4) can be a lot larger than their difference, and so you can get large errors in s^2 if you round off $(\sum_{i=1}^{n} x_i)^2$ and/or $\sum_{i=1}^{n}(x_i)^2$ (or, in extreme cases, a computer does it for you). See `http://www.johndcook.com/blog/2008/09/28/theoretical-explanation-for-numerical-results/`

0.3.2 Random Variables

The second concept to recall from your previous course is the concept of a random variable. We generally write this with a capital letter X, and define the probability it has to take on certain values. For any random variable, we can define the *cumulative distribution function* (cdf)

$$F(x) = P(X \le x) \tag{0.5}$$

If there are multiple random variables and we need to specify which one we're talking about, we may write this $F_X(x)$.

A *discrete random variable* can take one of a (possibly infinite) set of values, and the probability of it taking a particular value is given by the *probability mass function* (pmf, written $p_X(x)$ if necessary)

$$p(x) = P(X = x) \tag{0.6}$$

The probability of X taking on one of a set of values A is the sum of the pmf over the values in that set:

$$P(X \in A) = \sum_{x \in A} p(x) \tag{0.7}$$

As a special case, the sum of all the pmf values is equal to the probability that the random variable takes on *some* value, i.e.,

$$\sum_{x} p(x) = 1 \tag{0.8}$$

This is the *normalization* condition for the pmf.

An example of a discrete random variable is a binomial random variable; this describes the situation where we do a set of "Bernoulli trials", experiments which each have the same probability p of "success" and have no influence on each other. If we do n such trials, the number of successes is a random variable X with pmf

$$p(x) = b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \tag{0.9}$$

where $x! = 1 \times 2 \times \cdots \times (x-1) \times x$ is the factorial of x, so that

$$\binom{n}{x} = \frac{n \times (n-1) \times \cdots \times (n-x+1) \times (n-x)}{x \times (x-1) \times \cdots \times 2 \times 1} \tag{0.10}$$

See Chapter Three of Devore and `http://ccrg.rit.edu/~whelan/courses/2011_4wi_1016_351/notes03.pdf` for more details on discrete random variables.

A *continuous random variable* has zero probability of taking any precise numerical value, but its probability of falling in a range of interest is defined by its *probability density function* (pdf) $f(x)$ (or $f_X(x)$)

$$P(a < X < b) = \int_a^b f(x)\,dx \tag{0.11}$$

Note that since there's zero probability that X equals exactly a or b, it doesn't matter if we write $<$ or $\leq$ in the probability. The normalization condition for the pdf of a continuous random variable is

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)\,dx = 1 \tag{0.12}$$

The pdf is the derivative of the pdf, so

$$f(x) = F'(x) \qquad \text{and} \qquad F(x) = \int_{-\infty}^{x} f(y)\,dy \tag{0.13}$$

One common type of continuous random variable is that described by a *normal distribution* (also known as a Gaussian distribution), which has a pdf described by parameters μ (which may be positive, negative or zero and σ (which must be positive)

$$f(x) = f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{0.14}$$

its cumulative distribution function is

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{0.15}$$

where the function $\Phi(z)$ is defined as

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\,du \tag{0.16}$$

See Chapter Four of Devore and `http://ccrg.rit.edu/~whelan/courses/2011_4wi_1016_351/notes04.pdf` for more details on continuous random variables.

An important quantity which can be calculated from a probability distribution (pmf or pdf) is the expected value $E(X)$, which is defined as a weighted average value constructed from the pmf or pdf:

$$E(X) = \sum_x x\,p(x) \qquad \text{or} \qquad E(X) = \int_{-\infty}^{\infty} x\,f(x)\,dx \tag{0.17}$$

This also works for any function of the random variable:

$$E(h(X)) = \sum_x h(x)\,p(x) \quad \text{or} \quad E(h(X)) = \int_{-\infty}^{\infty} h(x)\,f(x)\,dx \tag{0.18}$$

We often write the expected value $E(X)$ as μ or μ_X and refer to it as the *mean* of the distribution. This is analogous to the sample mean $\overline{x}$ of descriptive statistics, but instead of averaging over a specific set of values in the dataset, it's averaging over a hypothetical repeated set of measurements. This is also sometimes called the *population mean*.

One application of the expected value is the *variance* $V(X) = E([X - \mu_X])$, which is sometimes written σ^2 or σ_X^2. This is the analogue of the sample variance s^2. We sometimes call σ_X^2 the

variance of the distribution associated with X, or the *population variance*.

Finally, the median of the distribution, $\widetilde{\mu}$ or $\widetilde{\mu}_X$ is defined indirectly as the value that the random variable has at least a 50% chance of lying on either side of:

$$P(X \leq \widetilde{\mu}) \geq \frac{1}{2} \leq P(X \geq \widetilde{\mu}) \tag{0.19}$$

For a discrete distribution, this has the simpler form

$$\int_{-\infty}^{\widetilde{\mu}} f(x)\,dx = \frac{1}{2} = \int_{\widetilde{\mu}}^{\infty} f(x)\,dx \tag{0.20}$$

0.3.3 Random Samples

Recall the concept of joint probability distributions for multiple random variables. For instance, if X_1, X_2, and X_3 are discrete random variables, we can write the joint pmf

$$p(x_1, x_2, x_3) = P([X_1 = x_1] \cap [X_2 = x_2] \cap [X_3 = x_3]) \tag{0.21}$$

I.e., the probability that X_1 takes the value x_1, *and* X_2 takes the value x_2, *and* X_3 takes the value x_3. Likewise, if X_1 and X_2 are continuous random variables, the joint pdf $f(x_1, x_2)$ can be used to construct probabilities like

$$P([a<X_1<b]\cap[c<X_2<d]) = \int_c^d \left(\int_a^b f(x_1, x_2)\,dx_1 \right) dx_2 \tag{0.22}$$

We say that X_1, X_2,... X_n are *independent* random variables if, for any possible values of x_1, x_2,... x_n, the joint pdf (taking the continuous case for concreteness) can be written

$X_1, X_2, \ldots X_n$ independent means

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)\, f_2(x_2) \cdots f_n(x_n) \tag{0.23}$$

A special case of this is when all of the functions f_1, f_2, ..., f_n are actually the same function; then we say the random variables are *independent and identically distributed* (iid):

$X_1, X_2, \ldots X_n$ iid means

$$f(x_1, x_2, \ldots x_n) = f(x_1) f(x_2) \cdots f(x_n) \tag{0.24}$$

We refer to this as a *sample* of size n from the distribution $f(x)$.

Given n random variables X_1, X_2, ... X_n, we refer to any function of the rvs as a *statistic*. By its nature, a statistic is itself a random variable. A number of useful statistics are created by combining the rvs in a sample using the same formulas that define descriptive statistics from a dataset. For example:

- The sample mean is $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$
- The sample variance is $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 s$
- The sample median $\widetilde{X}$ is a random variable defined by sorting the n values returned by the random variables in the sample and picking the one in the middle.

The linearity of the expected value can be used to work out the expected values of linear combinations of random variables. In particular, if

$$Y = a_1 X_1 + a_2 X_2 + \cdots a_n X_n = \sum_{i=1}^n a_i X_i \tag{0.25}$$

then

$$\begin{aligned} \mu_Y &= E(Y) = a_1 E(X_1) + a_2 E(X_2) + \cdots + a_n E(X_n) \\ &= a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n \end{aligned} \tag{0.26}$$

In the case of the variance (writing it for $n = 2$ for compactness,

$$V(a_1 X_1 + a_2 X_2) = {a_1}^2 V(X_1) + 2a_1 a_2 \operatorname{Cov}(X_1, X_2) + {a_2}^2 V(X_2) \tag{0.27}$$

where

$$\mathrm{Cov}(X_1, X_2) = E([X_1 - \mu_1][X_2 - \mu_2]) \tag{0.28}$$

is the *covariance* of the random variables X_1 and X_2. An important result shows that independent random variables have zero covariance,[2] so

if $X_1, \ldots, X_n$ independent,

$$V(Y) = {a_1}^2 V(X_1) + {a_2}^2 V(X_2) + \cdots + {a_n}^2 V(X_n) \tag{0.29}$$

Returning to the case of a random sample, the statistical properties of the sample mean $\overline{X}$ and S^2 are of interest, specifically, if the distribution has mean $\mu = E(X_i)$ and variance $\sigma^2 = V(X_i) = E([X_i - \mu]^2)$,

$$E(\overline{X}) = \mu \qquad \text{and} \qquad V(\overline{X}) = \frac{1}{n}\,\sigma^2 \tag{0.30}$$

One important result (shown in `http://ccrg.rit.edu/~whelan/courses/2011_4wi_1016_351/notes05.pdf` for example) is that

$$E\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = (n-1)\sigma^2 \ ; \tag{0.31}$$

this means that the sample variance S^2, defined with $n-1$ in the denominator, has an expectation value

$$E(S^2) = \sigma^2 \tag{0.32}$$

This is why the sample variance s^2 generated from a data set is usually given as $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Finally, note that the normal distribution has some interesting properties:

[2]The converse is not true; zero covariance does *not* imply independence.

1. Any statistic constructed as a linear combination of normally-distributed random variables is itself normally distributed
2. The sum (or the mean) of a large number of iid random variables of almost any distribution is approximately normally distributed. This is known as the Central Limit Theorem.

See Chapter Five of Devore and `http://ccrg.rit.edu/~whelan/courses/2011_4wi_1016_351/notes05.pdf` for more details on joint distributions and random samples.

Properties of Sums of Random Variables

$$T_o = \sum_{i=1}^{n} X_i \tag{0.33}$$

Property	When is it true?
$E(T_o) = \sum_{i=1}^{n} E(X_i)$	Always
$V(T_o) = \sum_{i=1}^{n} V(X_i)$	When $\{X_i\}$ independent
T_o normally distributed	Exact, when $\{X_i\}$ normally distributed
	Approximate, when $n \gtrsim 30$ (Central Limit Theorem)

Practice Problems

1.39, 1.51, 3.37, 3.39, 3.41, 3.43, 4.17, 4.29, 5.39, 5.45, 5.55, 5.65, 5.89

1 Fundamentals of Point Estimation

Most of the applications we'll be considering this semester are within the context of a *random sample* of n independent, identically distributed (iid) random variables $X_1, \ldots X_n$ drawn from a distribution $f(x;\theta)$ whose precise form depends on a parameter θ whose value is generally unknown. This means the joint distribution function for the sample will be

$$f(x_1, \ldots, x_n) = f(x_1;\theta)\, f(x_2;\theta) \cdots f(x_n;\theta) \tag{1.1}$$

We'll be considering procedures that allow us, one way or another, to say something about the unknown value θ once we've observed particular numerical values $x_1, \ldots, x_n$ for the random variables $X_1, \ldots X_n$. It is also possible for the distribution to depend on the values of multiple parameters $\theta_1, \theta_2, \ldots$. For instance, if our model specifies a normal distribution with unknown mean μ and standard deviation σ, the joint pdf for $X_1, \ldots X_n$ will be

$$f(x_1, \ldots, x_n) = f(x_1;\mu,\sigma)\, f(x_2;\mu,\sigma) \cdots f(x_n;\mu,\sigma) \tag{1.2}$$

The simplest thing we can do is come up with a *point estimate* $\hat{\theta}$, which is our best estimate for the value of the parameter θ given the data $x_1, \ldots, x_n$. For example, if the parameter θ is the mean value μ of the distribution,

$$\mu = E(X_i) = \int_{-\infty}^{\infty} x\, f(x;\mu)\, dx \tag{1.3}$$

one choice we can make for an estimate is the sample mean

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.4}$$

Note that there is a distinction between the estimate $\overline{x}$, which is a number derived from the actual observed data, and the *estimator* $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, which is a random variable (also known as a statistic) constructed out of the n random variables in the random sample. Statements we'll make about probability are actually about the estimator $\overline{X}$, for example, we can define the probability $P(\overline{X} > \mu)$, which is a statement about the behavior of the estimator, but something like $P(\overline{x} > \mu)$ can't actually be meaningfully defined in classical statistics, since $\overline{x}$ is a number we know from the data, and μ is a specific value, even if we know it; since neither $\overline{x}$ nor μ is random, this probability is either 0 or 1, depending on their values.[3] This distinction is somewhat obscured by the general notation, which relies on context to distinguish between the estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ and the estimate $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$.

1.1 Bias of an Estimator

If an estimator $\hat{\theta}$ were perfect, it would equal θ, so $\hat{\theta} - \theta$ is a random variable which describes the error made in trying to estimate θ. If the average value of that error is positive, it means in some sense the estimator overestimates θ, while if it's negative it underestimates it. This average error is called the bias of the estimator:

$$\left(\text{bias of } \hat{\theta}\right) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta = \mu_{\hat{\theta}} - \theta \tag{1.5}$$

An estimator for which $E(\hat{\theta}) = \theta$ is called *unbiased*.

For example, suppose $f(x;\mu)$ is any probability distribution with expectation value μ; then $\overline{X}$ is an unbiased estimator of μ:

[3] The situation is different in the field of Bayesian statistics, where probabilities can describe incomplete knowledge (in this case of μ) and not just randomness.

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu$$
$$= \frac{\overbrace{\mu + \cdots + \mu}^{n \text{ terms}}}{n} = \mu \tag{1.6}$$

Similarly, a result from MATH 251 (see for example section 3.4 of `http://ccrg.rit.edu/~whelan/courses/2011_4wi_1016_351/notes05.pdf`) shows that the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{1.7}$$

is an unbiased estimator of the variance σ^2 of the underlying probability distribution from which $X_1, \ldots, X_n$, since $E(S^2) = \sigma^2$. Note, though, that in general $E(\sqrt{S^2}) \neq \sigma$ (its actual value depends on the detailed shape of the probability distribution), so the sample standard deviation is not an unbiased estimator of the standard deviation of the probability distribution. (You will explore a related phenomenon on the homework.)

1.2 Variance of an Estimator

While it's often possible to find an unbiased estimator, so that the error $\hat{\theta} - \theta$ is zero on average, the squared error $(\hat{\theta} - \theta)^2$ will generally be positive, so we can consider the average squared error:

$$E\left([\hat{\theta} - \theta]^2\right) = E\left(\left[(\hat{\theta} - \mu_{\hat{\theta}}) + (\mu_{\hat{\theta}} - \theta)\right]^2\right)$$
$$= E\left((\hat{\theta} - \mu_{\hat{\theta}})^2\right) + 2(\mu_{\hat{\theta}} - \theta)\underbrace{E(\hat{\theta} - \mu_{\hat{\theta}})}_{\to 0} + (\mu_{\hat{\theta}} - \theta)^2$$
$$= V(\hat{\theta}) + (\mu_{\hat{\theta}} - \theta)^2 \tag{1.8}$$

This is the square of the bias of the estimator plus its variance (as a random variable). So it's reasonable that we'd like to pick an estimator with as small a variance as possible, especially if we limit ourselves to unbiased estimators. In particular, the minimum variance unbiased estimator (MVUE) is of some interest in classical statistics.

It's easy to see that there can be different unbiased estimators with different variances. Suppose again that the parameter of interest is the mean μ. We know the sample mean $\overline{X}$ is an unbiased estimator, but there are lots of ways to make unbiased estimator out of a linear combination of the random variables. For instance, just taking the first random variable X_1 as our estimator and throwing out the rest of the sample would also produced an unbiased estimator of μ, since $E(X_1) = \mu$. It's the variance that sets these choices apart. The variance of X_1 is just the variance of the underlying distribution, $V(X_1) = \sigma^2$. On the other hand, we know from MATH 251 that $V(\overline{X}) = \sigma^2/n$, so $\overline{X}$ is an estimator with lower variance than X_1.

The square root of the variance of an estimator is known as its *standard error*:

$$\left(\text{standard error of } \hat{\theta}\right) = \sqrt{V(\hat{\theta})} = \sigma_{\hat{\theta}} \tag{1.9}$$

Practice Problems

6.11, 6.13, 6.17, 6.19

Thursday 17 January 2019

2 Methods of Point Estimation

Last time we considered a random sample $X_1, \ldots, X_n$ drawn from a distribution described by one or more parameters (e.g., θ) and considered the properties of an estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ constructed from the sample. Now we consider two prescriptions for constructing estimators.

2.1 Method of Moments

Last time we used the example of the sample mean $\overline{X}$ as an estimator for the distribution mean μ. This method can be extended to cases where the parameter may not be simply realizable as a mean, e.g., the rate parameter λ of an exponential distribution $f(x;\lambda) = \lambda e^{-\lambda x}$. In this case, we consider the mean of the distribution

$$E(X) = \int_{-\infty}^{\infty} x\, f(x;\lambda)\, dx = \int_0^{\infty} x\lambda\, e^{-\lambda x}\, dx = \frac{1}{\lambda} \tag{2.1}$$

and define the estimate $\hat{\lambda}$ as the value which makes $E(X)$ equal the actual $\overline{x}$ observed in the data, i.e.,

$$\frac{1}{\hat{\lambda}} = \overline{x} \tag{2.2}$$

or $\hat{\lambda} = 1/\overline{x}$. The estimator is the corresponding quantity expressed in terms of the random sample, $\hat{\lambda} = 1/\overline{X}$.

This method can also be extended to situations where there are multiple parameters, using moments of the data and the probability distribution:

$$(k\text{th moment of data}) = \frac{1}{n}\sum_{i=1}^{n}(x_i)^k = \overline{x^k} \tag{2.3}$$

$$(k\text{th moment of distribution}) = \int_{-\infty}^{\infty} x^k\, f(x;\theta)\, dx = E[X^k] \tag{2.4}$$

For example, if the distribution is the normal distribution

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{2.5}$$

there are two parameters, so we need the first two moments of the distribution, which are

$$E(X) = \mu \tag{2.6a}$$

$$E(X^2) = V(X) + \left(E(X)\right)^2 = \sigma^2 + \mu^2 \tag{2.6b}$$

we therefore define the method-of-moments estimates $\hat{\mu}$ and $\hat{\sigma}$ by the equations

$$\overline{x} = \hat{\mu} \tag{2.7a}$$

$$\overline{x^2} = \hat{\sigma}^2 + \hat{\mu}^2 \tag{2.7b}$$

which we can solve for

$$\hat{\mu} = \overline{x} \tag{2.8a}$$

$$\hat{\sigma} = \sqrt{\overline{x^2} - \overline{x}^2} \tag{2.8b}$$

Note that the usual shortcut formula means that

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n-1}{n}\, s^2 \tag{2.9}$$

The estimators are then $\hat{\mu} = \overline{X}$ and $\hat{\sigma} = \sqrt{\frac{n-1}{n}\, S^2}$. Note that

$$E(\hat{\sigma}^2) = \frac{n-1}{n} E(S^2) = \frac{n-1}{n}\, \sigma^2 \tag{2.10}$$

So the method of moments estimator (in this case for σ^2) need not be unbiased.

2.2 Maximum Likelihood Estimation

A widely used method of parameter estimation starts with the likelihood function. This is actually the joint pdf for the data, evaluated at the actual data points:

$$f(x_1, \ldots, x_n; \theta) = f(x_1; \theta)\, f(x_2; \theta) \ldots f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{2.11}$$

we view this as a function of the parameter(s), and choose the parameter value(s) which make(s) it as large as possible.

To give a concrete example, consider a sample of size n drawn from an exponential distribution

$$f(x; \lambda) = \lambda\, e^{-\lambda x} \qquad 0 < \lambda < \infty,\ 0 < x < \infty \tag{2.12}$$

The likelihood function is

$$f(x_1, \ldots, x_n; \lambda) = \left(\lambda\, e^{-\lambda x_1}\right)\left(\lambda\, e^{-\lambda x_2}\right) \cdots \left(\lambda\, e^{-\lambda x_n}\right) = \lambda^n e^{-\lambda(x_1 + \cdots + x_n)} \tag{2.13}$$

Now, to find the λ which maximizes this for a particular $x_1, \ldots, x_n$, we need to take the derivative of the likelihood function and set it to zero. Since we're treating x_1, x_2, etc as constants, this is technically the *partial* derivative, and we can calculate it using the product rule:

$$\begin{aligned} &\frac{\partial}{\partial \lambda} f(x_1, \ldots, x_n; \lambda) \\ &\quad = n\lambda^{n-1}\, e^{-\lambda(x_1 + \cdots + x_n)} - \lambda^n (x_1 + \cdots + x_n) e^{-\lambda(x_1 + \cdots + x_n)} \\ &\quad = \left[n - \lambda(x_1 + \cdots + x_n)\right] \lambda^{n-1}\, e^{-\lambda(x_1 + \cdots + x_n)} \end{aligned} \tag{2.14}$$

Since λ^{n-1} and $e^{-\lambda(x_1 + \cdots + x_n)}$ are both guaranteed to be positive, the only way to get the derivative to be zero is for the quantity in square brackets to be zero, which gives us the maximum likelihood estimate

$$\hat{\lambda} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{\overline{x}} \tag{2.15}$$

For due diligence, we should also check that the maximum of the function is not at the boundary, in this case $\lambda = 0$. We can see that in fact $f(x_1, \ldots, x_n; 0) = 0$, so $\lambda = 0$ is actually the *minimum* of the likelihood function, and $\hat{\lambda}$ is indeed the maximum.

It actually turns out to be easier to take the logarithm[4] of the likelihood function first, and then take the derivative. And the parameter value which maximizes $\ln f(x_1, \ldots, x_n; \lambda)$ will also maximize $f(x_1, \ldots, x_n; \lambda)$, so it should give the same result. To see this in action, we note that for the exponential case,

$$\begin{aligned} \ln f(x_1, \ldots, x_n; \lambda) &= \ln\left(\lambda^n\, e^{-\lambda(x_1 + \cdots + x_n)}\right) \\ &= \ln\left(\lambda^n\right) + \ln\left(e^{-\lambda(x_1 + \cdots + x_n)}\right) = n \ln \lambda - \lambda(x_1 + \cdots + x_n) \end{aligned} \tag{2.16}$$

Now we can use the sum rule instead of the product rule:

$$\frac{\partial}{\partial \lambda} \ln f(x_1, \ldots, x_n; \lambda) = \frac{n}{\lambda} - (x_1 + \cdots + x_n) \tag{2.17}$$

Once again the solution is $\hat{\lambda} = 1/\overline{x}$, which makes the maximum likelihood estimator $\hat{\lambda} = 1/\overline{X}$.

2.3 Perspective

The maximum likelihood estimator (MLE) has a number of useful properties. It gives the same estimates for the parameters even if you change variables (say from variance to standard deviation, or from rate parameter λ to scale parameter $\beta = 1/\lambda$). And while the MLE is not necessarily unbiased, the bias goes to zero as the sample size gets large, and in fact in the limit of

[4] This is the natural logarithm, ln, i.e., the logarithm base e. Note that this will always be defined, since the likelihood function, being constructed from a pdf, is never negative.

large sample sizes, the MLE will approach the minimum variance unbiased estimator (MVUE) discussed in the last class.

It's tempting to thing of the maximum likelihood estimate as the most likely parameter value in light of the data, but that's not really correct. $f(\{x_i\};\theta)$ is a probability distribution for the data $\{x_i\}$, *not* for the parameter θ. In fact, classical statistics doesn't even let us define a probability distribution for the fixed but unknown value of the parameter θ. In Bayesian statistics, the rules are different, and one can define what's known as the *posterior probability distribution* $f(\theta|\{x_i\})$ for the parameter, given the observed data. This is defined by a version of Bayes's theorem (see Chapter Two of Devore); the likelihood $f(\{x_i\};\theta)$ can be thought of as a conditional distribution for a given value of θ, so we write it as $f(\{x_i\}|\theta)$ and the posterior distribution as

$$f(\theta|\{x_i\}) = \frac{f(\{x_i\}|\theta)\, f(\theta)}{f(\{x_i\})} \tag{2.18}$$

There's a whole field of statistical inference dealing with this equation and its consequences. The interesting point here is that, if the *prior probability distribution* $f(\theta)$ (which encodes our knowledge about the value of θ before drawing the sample $\{x_i\}$) is a constant, then the posterior $f(\theta|\{x_i\})$ is proportional to the likelihood $f(\{x_i\}|\theta)$, and so the θ value with the highest posterior probability (which actually *is* the most plausible value given the data) will in fact be the maximum likelihood value. So this is sort of a backdoor explanation of why the maximum-likelihood estimate is a reasonable thing.

Practice Problems

6.23, 6.25, 6.29, 6.37