Point Estimation
(Devore Chapter Six)

MATH-252-01: Probability and Statistics II*

Fall 2016

Contents

Tuesday 23 August 2016

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 8th edition, which is probably the version you used in MATH 251. There is a new 9th edition out, but the changes between editions should be minimal.
- 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. 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, and I'll be having some of my office hours in 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:

  9% Problem Sets & Computer Exercises
  6% 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 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: although we'll do a brief review this week, you will generally be expected to recall and apply what you learned in MATH 251.

1 Review of Statistical Formalism

1.1 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 $x_i \le \widetilde{x}$ and at least half have $x_i \ge \widetilde{x}$. 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 $x_i \le 6.77$, and at least 30% have $x_i \ge 6.77$.

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

  $$\overline{x} = \frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5) = \frac{1}{5}\sum_{i=1}^{5} x_i$$
  $$= \frac{4.92 + 5.35 + 6.77 + 8.74 + 5.08}{5} = \frac{30.86}{5} = 6.172 \tag{1.1}$$

  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

  $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{1.2}$$

  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 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 in^2.

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

[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.

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

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i)^2 - n(\overline{x})^2 \tag{1.3}$$

which can be used to calculate the sample variance as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i)^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} x_i\right)^2 \tag{1.4}$$

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 (1.4) 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/`

1.2 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)

$$F(x) = P(X \leq x) \tag{1.5}$$

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)

$$p(x) = P(X = x) \tag{1.6}$$

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:

$$P(X \in A) = \sum_{x \in A} p(x) \tag{1.7}$$

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.,

$$\sum_{x} p(x) = 1 \tag{1.8}$$

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

$$p(x) = b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \tag{1.9}$$

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

$$\binom{n}{x} = \frac{n \times (n-1) \times \cdots \times (n-x+1) \times (n-x)}{x \times (x-1) \times \cdots \times 2 \times 1} \tag{1.10}$$

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 $f_X(x)$)

$$P(a < X < b) = \int_a^b f(x)\,dx \tag{1.11}$$

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

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)\,dx = 1 \tag{1.12}$$

The pdf is the derivative of the pdf, so

$$f(x) = F'(x) \qquad \text{and} \qquad F(x) = \int_{-\infty}^{x} f(y)\,dy \tag{1.13}$$

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)

$$f(x) = f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{1.14}$$

its cumulative distribution function is

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{1.15}$$

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

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\,du \tag{1.16}$$

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:

$$E(X) = \sum_x x\,p(x) \qquad \text{or} \qquad E(X) = \int_{-\infty}^{\infty} x\,f(x)\,dx \tag{1.17}$$

This also works for any function of the random variable:

$$E(h(X)) = \sum_x h(x)\,p(x) \quad \text{or} \quad E(h(X)) = \int_{-\infty}^{\infty} h(x)\,f(x)\,dx \tag{1.18}$$

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:

$$P(X \leq \widetilde{\mu}) \geq \frac{1}{2} \leq P(X \geq \widetilde{\mu}) \tag{1.19}$$

For a discrete distribution, this has the simpler form

$$\int_{-\infty}^{\widetilde{\mu}} f(x)\,dx = \frac{1}{2} = \int_{\widetilde{\mu}}^{\infty} f(x)\,dx \tag{1.20}$$

Practice Problems

1.39, 1.51, 3.37, 3.39, 3.41, 3.43, 4.17

Thursday 25 August 2016

1.3 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

$$p(x_1, x_2, x_3) = P([X_1 = x_1] \cap [X_2 = x_2] \cap [X_3 = x_3]) \quad (1.21)$$

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

$$P([a<X_1<b]\cap[c<X_2<d]) = \int_c^d \left(\int_a^b f(x_1, x_2)\, dx_1 \right) dx_2 \quad (1.22)$$

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

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)\, f_2(x_2) \cdots f_n(x_n) \quad (1.23)$$

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

$$f(x_1, x_2, \ldots x_n) = f(x_1) f(x_2) \cdots f(x_n) \quad (1.24)$$

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

$$Y = a_1 X_1 + a_2 X_2 + \cdots a_n X_n = \sum_{i=1}^n a_i X_i \quad (1.25)$$

then

$$\begin{aligned} \mu_Y &= E(Y) = a_1 E(X_1) + a_2 E(X_2) + \cdots + a_n E(X_n) \\ &= a_1 \mu_1 + a_2 \mu_2 + \cdots + a_n \mu_n \end{aligned} \quad (1.26)$$

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

$$V(a_1 X_1 + a_2 X_2) = {a_1}^2 V(X_1) + 2 a_1 a_2 \operatorname{Cov}(X_1, X_2) + {a_2}^2 V(X_2) \quad (1.27)$$

where

$$\operatorname{Cov}(X_1, X_2) = E([X_1 - \mu_1][X_2 - \mu_2]) \quad (1.28)$$

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,

$$V(Y) = {a_1}^2 V(X_1) + {a_2}^2 V(X_2) + \cdots + {a_n}^2 V(X_n) \quad (1.29)$$

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 $\sigma^2 = V(X_i) = E([X_i - \mu]^2)$,

$$E(\overline{X}) = \mu \qquad \text{and} \qquad V(\overline{X}) = \frac{1}{n}\,\sigma^2 \quad (1.30)$$

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

$$E\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = (n-1)\sigma^2 \; ; \quad (1.31)$$

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

$$E(S^2) = \sigma^2 \quad (1.32)$$

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:

1. Any statistic constructed as a linear combination of normally-distributed random variables is itself normally distributed

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

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.

1.3.1 Properties of Sums of Random Variables

$$T_o = \sum_{i=1}^{n} X_i \quad (1.33)$$

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

4.29, 5.39, 5.45, 5.55, 5.65, 5.89

Tuesday 30 August 2016

Guest lecture by Dr. Richard O'Shaughnessy

2 Fundamentals of Point Estimation

Practice Problem

6.11, 6.13, 6.17, 6.19

Thursday 1 September 2016

Guest lecture by Dr. Richard O'Shaughnessy

3 Methods of Point Estimation

3.1 Method of Moments

3.2 Maximum Likelihood Estimation

Practice Problems

6.23, 6.25, 6.29, 6.37