

Introduction to Statistical Thought

Michael Lavine

December 18, 2006

CONTENTS

List of Figures	vii
List of Tables	xi
Preface	xiii
1 Probability	1
1.1 Basic Probability	1
1.2 Probability Densities	6
1.3 Parametric Families of Distributions	14
1.3.1 The Binomial Distribution	15
1.3.2 The Poisson Distribution	18
1.3.3 The Exponential Distribution	21
1.3.4 The Normal Distribution	24
1.4 Centers, Spreads, Means, and Moments	30
1.5 Joint, Marginal and Conditional Probability	42
1.6 Association, Dependence, Independence	53
1.7 Simulation	60
1.7.1 Calculating Probabilities	60
1.7.2 Evaluating Statistical Procedures	65
1.8 R	76
1.9 Some Results for Large Samples	81
1.10 Exercises	87
2 Modes of Inference	99
2.1 Data	99
2.2 Data Description	100

2.2.1	Summary Statistics	101
2.2.2	Displaying Distributions	106
2.2.3	Exploring Relationships	120
2.3	Likelihood	134
2.3.1	The Likelihood Function	134
2.3.2	Likelihoods from the Central Limit Theorem	144
2.3.3	Likelihoods for several parameters	148
2.4	Estimation	158
2.4.1	The Maximum Likelihood Estimate	158
2.4.2	Accuracy of Estimation	160
2.4.3	The sampling distribution of an estimator	163
2.5	Bayesian Inference	169
2.6	Prediction	180
2.7	Hypothesis Testing	184
2.8	Exercises	198
3	Regression	209
3.1	Introduction	209
3.2	Normal Linear Models	218
3.2.1	Introduction	218
3.2.2	Inference for Linear Models	230
3.3	Generalized Linear Models	244
3.3.1	Logistic Regression	244
3.3.2	Poisson Regression	253
3.4	Predictions from Regression	257
3.5	Exercises	261
4	More Probability	273
4.1	More Probability Density	273
4.2	Random Vectors	274
4.2.1	Densities of Random Vectors	275
4.2.2	Moments of Random Vectors	276
4.2.3	Functions of Random Vectors	276
4.3	Representing Distributions	281
4.4	Exercises	286

5	Special Distributions	289
5.1	Binomial and Negative Binomial	289
5.2	Multinomial	300
5.3	Poisson	302
5.4	Uniform	310
5.5	Gamma, Exponential, Chi Square	310
5.6	Beta	318
5.7	Normal	319
5.7.1	The Univariate Normal Distribution	319
5.7.2	The Multivariate Normal Distribution	325
5.8	t and F	334
5.8.1	The t distribution	334
5.8.2	The F distribution	339
5.9	Exercises	339
6	More Models	349
6.1	Hierarchical Models	349
6.2	Time Series and Markov Chains	350
6.3	Contingency Tables	364
6.4	Survival analysis	364
6.5	The Poisson process	364
6.6	Change point models	364
6.7	Evaluating and enhancing models	365
6.8	Exercises	365
7	Mathematical Statistics	367
7.1	Properties of Statistics	367
7.1.1	Sufficiency	367
7.1.2	Consistency, Bias, and Mean-squared Error	370
7.1.3	Efficiency	372
7.1.4	Asymptotic Normality	372
7.1.5	Robustness	372
7.2	Transformations of Parameters	372
7.3	Information	373
7.4	More Hypothesis Testing	373
7.4.1	p values	373
7.4.2	The Likelihood Ratio Test	373
7.4.3	The Chi Square Test	373

7.4.4	Power	373
7.5	Exponential families	373
7.6	Location and Scale Families	373
7.7	Functionals	373
7.8	Invariance	373
7.9	Asymptotics	373
7.10	Exercises	380
Bibliography		383
Index		385
Index of Examples		389

LIST OF FIGURES

1.1	pdf for time on hold at Help Line	7
1.2	p_Y for the outcome of a spinner	9
1.3	(a) : Ocean temperatures; (b) : Important discoveries	11
1.4	Change of variables	14
1.5	Binomial probabilities	17
1.6	$P[X = 3 \lambda]$ as a function of λ	20
1.7	Exponential densities	22
1.8	Normal densities	25
1.9	Ocean temperatures at 45°N, 30°W, 1000m depth	27
1.10	Normal samples and Normal densities	29
1.11	hydrographic stations off the coast of Europe and Africa	32
1.12	Water temperatures	33
1.13	Two pdf's with ± 1 and ± 2 SD's.	39
1.14	Water temperatures with standard deviations	43
1.15	Permissible values of N and X	47
1.16	Features of the joint distribution of (X, Y)	51
1.17	Lengths and widths of sepals and petals of 150 iris plants	55
1.18	correlations	58
1.19	1000 simulations of $\hat{\theta}$ for <code>n.sim</code> = 50, 200, 1000	64
1.20	1000 simulations of $\hat{\theta}$ under three procedures	68
1.21	Monthly concentrations of CO ₂ at Mauna Loa	69
1.22	1000 simulations of a FACE experiment	73
1.23	Histograms of craps simulations	86
2.1	quantiles	103
2.2	Histograms of tooth growth	108
2.3	Histograms of tooth growth	109

2.4	Histograms of tooth growth	110
2.5	calorie contents of beef hot dogs	114
2.6	Strip chart of tooth growth	116
2.7	Quiz scores from Statistics 103	119
2.8	QQ plots of water temperatures ($^{\circ}\text{C}$) at 1000m depth	121
2.9	Mosaic plot of <code>UCBAdmissions</code>	125
2.10	Mosaic plot of <code>UCBAdmissions</code>	126
2.11	Old Faithful data.	129
2.12	Waiting time versus duration in the Old Faithful dataset	130
2.13	Time series of duration and waiting time at Old Faithful	131
2.14	Time series of duration and waiting time at Old Faithful	132
2.15	Temperature versus latitude for different values of longitude	135
2.16	Temperature versus longitude for different values of latitude	136
2.17	Likelihood function for the proportion of red cars	138
2.18	$\ell(\theta)$ after $\sum y_i = 40$ in 60 quadrats.	141
2.19	Likelihood for Slater School	142
2.20	Marginal and exact likelihoods for Slater School	145
2.21	Marginal likelihood for mean CEO salary	147
2.22	FACE Experiment: data and likelihood	150
2.23	Likelihood function for Quiz Scores	153
2.24	Log of the likelihood function for (λ, θ_f) in Example 2.12	157
2.25	Likelihood function for the probability of winning craps	162
2.26	Sampling distribution of the sample mean and median	165
2.27	Histograms of the sample mean for samples from $\text{Bin}(n, .1)$	167
2.28	Prior, likelihood and posterior in the seedlings example	174
2.29	Prior, likelihood and posterior densities for λ with $n = 1, 4, 16$	176
2.30	Prior, likelihood and posterior densities for λ with $n = 60$	177
2.31	Prior, likelihood and posterior density for Slater School	179
2.32	Plug-in predictive distribution for seedlings	181
2.33	Predictive distributions for seedlings after $n = 0, 1, 60$	185
2.34	pdf of the $\text{Bin}(100, .5)$ distribution	190
2.35	pdfs of the $\text{Bin}(100, .5)$ (dots) and $\text{N}(50, 5)$ (line) distributions	191
2.36	Approximate density of summary statistic t	193
2.37	Number of times baboon father helps own child	197
2.38	Histogram of simulated values of <code>w.tot</code>	198
3.1	Four regression examples	211
3.2	1970 draft lottery. Draft number vs. day of year	213

3.3	Draft number vs. day of year with smoothers	214
3.4	Total number of New seedlings 1993 – 1997, by quadrat.	216
3.5	Calorie content of hot dogs	219
3.6	Density estimates of calorie contents of hot dogs	221
3.7	The <code>PlantGrowth</code> data	223
3.8	Ice cream consumption versus mean temperature	229
3.9	Likelihood functions for $(\mu, \delta_M, \delta_P)$ in the Hot Dog example.	236
3.10	<code>pairs</code> plot of the <code>mtcars</code> data	238
3.11	<code>mtcars</code> — various plots	241
3.12	likelihood functions for $\beta_1, \gamma_1, \delta_1$ and δ_2 in the <code>mtcars</code> example.	243
3.13	Pine cones and O-rings	246
3.14	Pine cones and O-rings with regression curves	248
3.15	Likelihood function for the pine cone data	250
3.16	Actual vs. fitted and residuals vs. fitted for the seedling data	256
3.17	Diagnostic plots for the seedling data	258
3.18	Actual mpg and fitted values from three models	260
3.19	Happiness Quotient of bankers and poets	264
4.1	The (X_1, X_2) plane and the (Y_1, Y_2) plane	280
4.2	pmf's, pdf's, and cdf's	282
5.1	The Binomial pmf	295
5.2	The Negative Binomial pmf	299
5.3	Poisson pmf for $\lambda = 1, 4, 16, 64$	305
5.4	Rutherford and Geiger's Figure 1	309
5.5	Gamma densities	312
5.6	Exponential densities	315
5.7	Beta densities	320
5.8	Water temperatures ($^{\circ}\text{C}$) at 1000m depth	321
5.9	Bivariate Normal density	328
5.10	Bivariate Normal density	331
5.11	t densities for four degrees of freedom and the $N(0, 1)$ density	338
6.1	Graphical representation of hierarchical model for fMRI	350
6.2	Some time series	352
6.3	Y_{t+1} vs. Y_t for the Beaver and Presidents data sets	354
6.4	Y_{t+k} vs. Y_t for the Beaver data set and lags 0–5	355
6.5	coplot of $Y_{t+1} \sim Y_{t-1} Y_t$ for the Beaver data set	357

6.6	Fit of CO ₂ data	360
6.7	DAX closing prices	361
6.8	DAX returns	363
7.1	The Be(.39, .01) density	377
7.2	Densities of $\bar{Y}_{in}$	379
7.3	Densities of Z_{in}	381

LIST OF TABLES

1.1	Party Affiliation and Referendum Support	44
1.2	Steroid Use and Test Results	46
2.1	New and Old seedlings in quadrat 6 in 1992 and 1993	155
3.1	Correspondence between Models 3.3 and 3.4	222
3.2	β 's for Figure 3.14	247
5.1	Rutherford and Geiger's data	308

PREFACE

This book is intended as an upper level undergraduate or introductory graduate textbook in statistical thinking with a likelihood emphasis for students with a good knowledge of calculus and the ability to think abstractly. By “statistical thinking” is meant a focus on ideas that statisticians care about as opposed to technical details of how to put those ideas into practice. By “likelihood emphasis” is meant that the likelihood function and likelihood principle are unifying ideas throughout the text. Another unusual aspect is the use of statistical software as a pedagogical tool. That is, instead of viewing the computer merely as a convenient and accurate calculating device, we use computer calculation and simulation as another way of explaining and helping readers understand the underlying concepts.

Our software of choice is **R**. **R** and accompanying manuals are available for free download from <http://www.r-project.org>. You may wish to download **An Introduction to R** to keep as a reference. It is highly recommended that you try all the examples in **R**. They will help you understand concepts, give you a little programming experience, and give you facility with a very flexible statistical software package. And don’t just try the examples as written. Vary them a little; play around with them; experiment. You won’t hurt anything and you’ll learn a lot.

CHAPTER 1

PROBABILITY

1.1 Basic Probability

Let $\mathcal{X}$ be a set and $\mathcal{F}$ a collection of subsets of $\mathcal{X}$. A *probability measure*, or just a *probability*, on $(\mathcal{X}, \mathcal{F})$ is a function $\mu : \mathcal{F} \rightarrow [0, 1]$. In other words, to every set in $\mathcal{F}$, μ assigns a probability between 0 and 1. We call μ a *set function* because its domain is a collection of sets. But not just any set function will do. To be a probability μ must satisfy

1. $\mu(\emptyset) = 0$ ($\emptyset$ is the empty set.),
2. $\mu(\mathcal{X}) = 1$, and
3. if A_1 and A_2 are disjoint then $\mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2)$.

One can show that property 3 holds for any finite collection of disjoint sets, not just two; see Exercise 1. It is common practice, which we adopt in this text, to assume more — that property 3 also holds for any countable collection of disjoint sets.

When $\mathcal{X}$ is a finite or countably infinite set (usually integers) then μ is said to be a *discrete* probability. When $\mathcal{X}$ is an interval, either finite or infinite, then μ is said to be a *continuous* probability. In the discrete case, $\mathcal{F}$ usually contains all possible subsets of $\mathcal{X}$. But in the continuous case, technical complications prohibit $\mathcal{F}$ from containing all possible subsets of $\mathcal{X}$. See Casella and Berger [2002] or Schervish [1995] for details. In this text we deemphasize the role of $\mathcal{F}$ and speak of probability measures on $\mathcal{X}$ without mentioning $\mathcal{F}$.

In practical examples $\mathcal{X}$ is the set of outcomes of an “experiment” and μ is determined by experience, logic or judgement. For example, consider rolling a six-sided die. The set of outcomes is $\{1, 2, 3, 4, 5, 6\}$ so we would assign $\mathcal{X} \equiv \{1, 2, 3, 4, 5, 6\}$. If we believe the die to be fair then we would also assign $\mu(\{1\}) = \mu(\{2\}) = \cdots = \mu(\{6\}) = 1/6$. The laws of probability then imply various other values such as

$$\begin{aligned}\mu(\{1, 2\}) &= 1/3 \\ \mu(\{2, 4, 6\}) &= 1/2 \\ \text{etc.}\end{aligned}$$

Often we omit the braces and write $\mu(2)$, $\mu(5)$, etc. Setting $\mu(i) = 1/6$ is not automatic simply because a die has six faces. We set $\mu(i) = 1/6$ because we believe the die to be fair.

We usually use the word “probability” or the symbol P in place of μ . For example, we would use the following phrases interchangeably:

- The probability that the die lands 1
- $P(1)$
- $P[\text{the die lands 1}]$
- $\mu(\{1\})$

We also use the word *distribution* in place of *probability measure*.

The next example illustrates how probabilities of complicated events can be calculated from probabilities of simple events.

Example 1.1 (The Game of Craps)

Craps is a gambling game played with two dice. Here are the rules, as explained on the website www.online-craps-gambling.com/craps-rules.html.

For the dice thrower (shooter) the object of the game is to throw a 7 or an 11 on the first roll (a win) and avoid throwing a 2, 3 or 12 (a loss). If none of these numbers (2, 3, 7, 11 or 12) is thrown on the first throw (the Come-out roll) then a Point is established (the point is the number rolled) against which the shooter plays. The shooter continues to throw until one of two numbers is thrown, the Point number or a Seven. If the shooter rolls the Point before rolling a Seven he/she wins, however if the shooter throws a Seven before rolling the Point he/she loses.

Ultimately we would like to calculate $P(\text{shooter wins})$. But for now, let's just calculate

$$P(\text{shooter wins on Come-out roll}) = P(7 \text{ or } 11) = P(7) + P(11).$$

Using the language of page 1, what is $\mathcal{X}$ in this case? Let d_1 denote the number showing on the first die and d_2 denote the number showing on the second die. d_1 and d_2 are integers from 1 to 6. So $\mathcal{X}$ is the set of ordered pairs (d_1, d_2) or

$$\begin{array}{cccccc} (6, 6) & (6, 5) & (6, 4) & (6, 3) & (6, 2) & (6, 1) \\ (5, 6) & (5, 5) & (5, 4) & (5, 3) & (5, 2) & (5, 1) \\ (4, 6) & (4, 5) & (4, 4) & (4, 3) & (4, 2) & (4, 1) \\ (3, 6) & (3, 5) & (3, 4) & (3, 3) & (3, 2) & (3, 1) \\ (2, 6) & (2, 5) & (2, 4) & (2, 3) & (2, 2) & (2, 1) \\ (1, 6) & (1, 5) & (1, 4) & (1, 3) & (1, 2) & (1, 1) \end{array}$$

If the dice are fair, then the pairs are all equally likely. Since there are 36 of them, we assign $P(d_1, d_2) = 1/36$ for any combination (d_1, d_2) . Finally, we can calculate

$$\begin{aligned} P(7 \text{ or } 11) &= P(6, 5) + P(5, 6) + P(6, 1) + P(5, 2) \\ &+ P(4, 3) + P(3, 4) + P(2, 5) + P(1, 6) = 8/36 = 2/9. \end{aligned}$$

The previous calculation uses desideratum 3 for probability measures. The different pairs $(6, 5)$, $(5, 6)$, $\dots$, $(1, 6)$ are disjoint, so the probability of their union is the sum of their probabilities.

Example 1.1 illustrates a common situation. We know the probabilities of some simple events like the rolls of individual dice, and want to calculate the probabilities of more complicated events like the success of a Come-out roll. Sometimes those probabilities can be calculated mathematically as in the example. Other times it is more convenient to calculate them by computer simulation. We frequently use R to calculate probabilities. To illustrate, Example 1.2 uses R to calculate by simulation the same probability we found directly in Example 1.1.

Example 1.2 (Craps, continued)

To simulate the game of craps, we will have to simulate rolling dice. That's like randomly sampling an integer from 1 to 6. The `sample()` command in R can do that. For example, the following snippet of code generates one roll from a fair, six-sided die and shows R's response:

```
> sample(1:6,1)
[1] 1
>
```

When you start R on your computer, you see `>`, R's prompt. Then you can type a command such as `sample(1:6,1)` which means "take a sample of size 1 from the numbers 1 through 6". (It could have been abbreviated `sample(6,1)`.) R responds with `[1] 1`. The `[1]` says how many calculations R has done; you can ignore it. The 1 is R's answer to the `sample` command; it selected the number "1". Then it gave another `>`, showing that it's ready for another command. Try this several times; you shouldn't get "1" every time.

Here's a longer snippet that does something more useful.

```
> x <- sample ( 6, 10, replace=T ) # take a sample of
                                   # size 10 and call it x

> x # print the ten values
[1] 6 4 2 3 4 4 3 6 6 2

> sum ( x == 3 ) # how many are equal to 3?
[1] 2
>
```

Note

- `#` is the comment character. On each line, R ignores all text after `#`.
- We have to tell R to take its sample *with replacement*. Otherwise, when R selects "6" the first time, "6" is no longer available to be sampled a second time. In `replace=T`, the T stands for True.
- `<-` does *assignment*. I.e., the result of `sample ( 6, 10, replace=T )` is assigned to a variable called `x`. The assignment symbol is two characters: `<` followed by `-`.
- A variable such as `x` can hold many values simultaneously. When it does, it's called a *vector*. You can refer to individual elements of a vector. For example, `x[1]` is the first element of `x`. `x[1]` turned out to be 6; `x[2]` turned out to be 4; and so on.

- `==` does *comparison*. In the snippet above, `(x==3)` checks, for each element of `x`, whether that element is equal to 3. If you just type `x == 3` you will see a string of T's and F's (True and False), one for each element of `x`. Try it.
- The `sum` command treats T as 1 and F as 0.
- R is almost always tolerant of spaces. You can often leave them out or add extras where you like.

On average, we expect 1/6 of the draws to equal 1, another 1/6 to equal 2, and so on. The following snippet is a quick demonstration. We simulate 6000 rolls of a die and expect about 1000 1's, 1000 2's, etc. We count how many we actually get. This snippet also introduces the `for` loop, which you should try to understand now because it will be *extremely* useful in the future.

```
> x <- sample(6,6000,replace=T)

> for ( i in 1:6 ) print ( sum ( x==i ))
[1] 995
[1] 1047
[1] 986
[1] 1033
[1] 975
[1] 964
>
```

Each number from 1 through 6 was chosen about 1000 times, plus or minus a little bit due to chance variation.

Now let's get back to craps. We want to simulate a large number of games, say 1000. For each game, we record either 1 or 0, according to whether the shooter wins on the Come-out roll, or not. We should print out the number of wins at the end. So we start with a code snippet like this:

```
# make a vector of length 1000, filled with 0's
wins <- rep ( 0, 1000 )
for ( i in 1:1000 ) {
```

```

    simulate a Come-out roll
    if shooter wins on Come-out, wins[i] <- 1
  }

sum ( wins ) # print the number of wins

```

Now we have to figure out how to simulate the Come-out roll and decide whether the shooter wins. Clearly, we begin by simulating the roll of two dice. So our snippet expands to

```

# make a vector of length 1000, filled with 0's
wins <- rep ( 0, 1000 )
for ( i in 1:1000 ) {
  d <- sample ( 1:6, 2, replace=T )
  if ( sum(d) == 7 || sum(d) == 11 ) wins[i] <- 1
}
sum ( wins ) # print the number of wins

```

The “||” stands for “or”. So that line of code sets `wins[i] <- 1` if the sum of the rolls is either 7 or 11. When I ran this simulation R printed out 219. The calculation in Example 1.1 says we should expect around $(2/9) \times 1000 \approx 222$ wins. Our calculation and simulation agree about as well as can be expected from a simulation. Try it yourself a few times. You shouldn't always get 219. But you should get around 222 plus or minus a little bit due to the randomness of the simulation.

Try out these R commands in the version of R installed on your computer. Make sure you understand them. If you don't, print out the results. Try variations. Try any tricks you can think of to help you learn R.

1.2 Probability Densities

So far we have dealt with *discrete* probabilities, or the probabilities of at most a countably infinite number of outcomes. For discrete probabilities, $\mathcal{X}$ is usually a set of integers, either finite or infinite. Section 1.2 deals with the case where $\mathcal{X}$ is an interval, either of finite or infinite length. Some examples are