

Notes on Statistics  
From a Bayesian Perspective

(last update May 9, 2007)

Brian S. Blais

© Copyright 2005  
by  
Brian S. Blais

# Preface: Why Make These Notes?

I'm writing these notes to clarify statistics for myself, never having had a formal class in statistics. It turns out that there is an interesting reason why I, as a physicist, have never taken a statistics course. This reason is best described in such references as (Loredo, 1990) and (Jaynes, 2003).

I have been hearing about “p-values” and “t-tests” for many years, and had only a vague understanding of them. I understood probability theory reasonably well, but somehow these terms that I was hearing didn't quite fit into the way I was thinking. After reading some recent work on Bayesian approaches, I realized that the standard (orthodox) approach to statistics made *no sense* to me, and the Bayesian way was entirely intuitive. Further, it seemed as if there were serious problems with some of the orthodox approaches, even on some trivial problems (see (Lindley and Phillips, 1976), and Section 4.1 for an illustrative example).

Still, many of the Bayesian articles and books that I read went over examples that were never covered in standard statistics books, and so their similarities and differences were not always clear. So I decided I would pick up a standard statistics book (in this case, (Bowerman and O'Connell, 2003)) and go through the examples and exercises in the book, but from a Bayesian perspective. Along the way, I leaned on such sources as (Jeffreys, 1939; Sivia, 1996; Jaynes, 2003), as well as numerous other articles, to get me through the Bayesian approach. These notes are the (work-in-progress) results.

I am also including Python ([www.python.org](http://www.python.org)) code for the examples, so I invite any readers to reproduce anything I've done.

## A Suggestion

If you are familiar with the orthodox perspectives, and are dubious of the Bayesian methods, I urge you to read in its entirety (Jaynes, 1976) (<http://bayes.wustl.edu/etj/articles/confidence.pdf>). It presents in detail six easy problems for which it is clear that the orthodox methods are lacking, and there are two responses from orthodox statisticians along with replies to those. It's an excellent read.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	History . . . . .	3
1.1.1	Derivation of Bayes' Theorem . . . . .	3
1.1.2	Further History . . . . .	3
1.1.3	Response . . . . .	4
1.2	Procedure . . . . .	5
1.2.1	Orthodox Hypothesis Tests . . . . .	6
1.3	Numerical Analysis . . . . .	6
1.3.1	Plotting Distributions and Histograms . . . . .	6
<b>2</b>	<b>One Sample Inferences</b>	<b>9</b>
2.1	Unknown $\mu$ , Known $\sigma$ . . . . .	9
2.1.1	Exercises . . . . .	10
2.2	Unknown $\mu$ , Unknown $\sigma$ . . . . .	15
2.2.1	Exercises . . . . .	17
2.3	Unknown proportion . . . . .	19
2.3.1	Confidence . . . . .	20
2.3.2	Median and Percentiles . . . . .	21
2.3.3	Numerical Examples . . . . .	22
<b>3</b>	<b>Two Sample Inferences</b>	<b>25</b>
3.1	Paired Data Difference of Means, $\delta_k \equiv x_k - y_k$ . . . . .	25
3.2	Difference of Means, $\delta \equiv \mu_x - \mu_y$ , known $\sigma_x$ and $\sigma_y$ . . . . .	25
3.3	Difference of Means, $\delta \equiv \mu_x - \mu_y$ , unknown $\sigma_x$ and $\sigma_y$ . . . . .	25
3.3.1	Jaynes 1976 Difference of Means . . . . .	29
3.4	Ratio of Two Variances $\kappa \equiv \sigma_x^2/\sigma_y^2$ . . . . .	31
3.5	Simple Linear Regression, $y_k = mx_k + b + \epsilon$ . . . . .	31
3.6	Linear Regression with Errors on both $x$ and $y$ . . . . .	32
3.7	Goodness of Fit . . . . .	33
3.7.1	Jaynes' Alternative to $\chi^2$ . . . . .	34

<b>4</b>	<b>Orthodox versus Bayesian Approaches</b>	<b>35</b>
4.1	Flipping a Tack . . . . .	35
4.1.1	Orthodox Statistics . . . . .	35
4.1.2	Bayesian Statistics . . . . .	37
4.2	Type A Stars . . . . .	37
4.3	Cauchy Distribution . . . . .	39
4.3.1	Orthodox estimator? . . . . .	40
<b>5</b>	<b>Misc</b>	<b>41</b>
5.1	Max Entropy Derivations of Priors . . . . .	41
5.1.1	Mean . . . . .	41
5.1.2	Mean and Second Moment . . . . .	41
5.2	Derivation of Maximum Entropy . . . . .	42
5.3	Problem from Loredo . . . . .	44
5.3.1	Estimating the Amplitude of a Signal . . . . .	44
5.3.2	Measuring a Weak Counting Signal . . . . .	45
5.4	Anova and T-distributions . . . . .	46
<b>A</b>	<b>Supplementary Code</b>	<b>50</b>
A.1	<code>utils.py</code> . . . . .	50
<b>B</b>	<b>Derivations for Single Samples</b>	<b>52</b>
B.1	Unknown $\mu$ , Known $\sigma$ . . . . .	52
B.1.1	Introducing the Distributions . . . . .	52
B.1.2	An Aside on Log Posterior . . . . .	53
B.1.3	Continuing . . . . .	53
B.2	Unknown $\mu$ , Unknown $\sigma$ . . . . .	54
B.2.1	Setting up the Problem . . . . .	55
B.2.2	Estimating the Mean . . . . .	56
B.2.3	A More Convenient Form . . . . .	57
B.2.4	Estimating $\sigma$ . . . . .	57
B.3	Unknown proportion . . . . .	59
B.3.1	Max, Mean, Variance . . . . .	60
<b>C</b>	<b>Derivations for Two Samples</b>	<b>62</b>
C.1	Paired Data Difference of Means, $\delta_k \equiv x_k - y_k$ . . . . .	62
C.1.1	Changing Variables . . . . .	62
C.1.2	Continuing with Paired Data . . . . .	63
C.2	Difference of Means, $\delta \equiv \mu_x - \mu_y$ , known $\sigma_x$ and $\sigma_y$ . . . . .	63

C.3	Difference of Means, $\delta \equiv \mu_x - \mu_y$ , unknown $\sigma_x$ and $\sigma_y$ . . . . .	65
C.4	Ratio of Two Variances $\kappa \equiv \sigma_x^2/\sigma_y^2$ . . . . .	65
C.5	Simple Linear Regression, $y_k = mx_k + b + \epsilon$ . . . . .	67
C.5.1	Quick recipe for solving $2 \times 2$ equations . . . . .	68
C.5.2	Solution to the Simple Least Squares Problem . . . . .	69
<b>D</b>	<b>Data</b> . . . . .	<b>70</b>
D.1	Bank Waiting Times (in Minutes) . . . . .	70
D.2	Customer Satisfaction (7-pt Likert Scale $\times$ 7 responses) . . . . .	70
<b>E</b>	<b>Probability Distributions and Integrals</b> . . . . .	<b>72</b>
E.1	Binomial . . . . .	72
E.1.1	Normalization . . . . .	72
E.1.2	Mean . . . . .	72
E.1.3	Variance . . . . .	73
E.1.4	Gaussian Approximation . . . . .	73
E.2	Negative Binomial . . . . .	74
E.3	Beta . . . . .	74
E.4	Gaussian . . . . .	74
E.4.1	Aside about Cool Tricks for Gaussian Integrals . . . . .	75

# Chapter 1

## Introduction

Probability is often defined as the “*long-run relative frequency of occurrence of an event*, either in a sequence of repeated experiments or in an ensemble of “identically prepared” systems.” (Loredo, 1990) This definition is referred to as the “frequentist” view, or the “orthodox” view of probability. The definition of probability used in these notes is the Bayesian one, where “probability is regarded as the real-number valued measure of the plausibility of a proposition when incomplete knowledge does not allow us to establish its truth or falsehood with certainty” (Loredo, 1990) In this way, it is a real-number extension of Boolean logic, where 0 represents certainty of falsehood and 1 represents certainty of truth.

Although this may seem like an argument for philosophers, there are real and demonstrable differences between the Bayesian and orthodox methods. In many cases, however, they lead to identical predictions. Most of the results in an introductory statistics text are derivable with either perspective. In some cases, the difference reflects different choices of problems under consideration.

Regarding Brian’s comments:

```
> > ...
> >I would refer everyone on this thread to the wonderful article "From
> >Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics"
> >by Tom Loredo, located at
> >http://www.astro.cornell.edu/staff/loredo/bayes/tjl.html
> >
> > From my personal experience trying to learn basic statistics, I
> >always got hung up on the notion of a population, and of the
> >standard deviation of the mean. I found the Bayesian approach to
> >be both more intuitive, easier to apply to real data, and more
> >mathematically sound (there is a great article by E.T. Jaynes at
> >http://bayes.wustl.edu/etj/articles/confidence.pdf where he outlines
> >several pathologies in standard stats).
```

I too, want to second Brian’s endorsement of the Bayesian approach to probability theory--especially as interpreted by Jaynes and his school of the maximum entropy procedure for determining probability

distributions based on incomplete data.

> >Bottom line: there is no population in the Bayesian approach.

True.

> >Probability is a measure of ones state of knowledge, not a property  
> >of the system.

Whoa, there. It's true that this approach does not interpret the fundamental meaning of probabilities as the asymptotic relative frequencies of particular outcomes for infinite ensembles of copies of the system (or random process) at hand. But a Bayesian interpretation of probability is not really based on anything as subjective as the state of knowledge any particular person's mind. Such a characterization is really a straw man that opponents of that approach tend to level at it. Rather, this interpretation of probability is just as objectively defined as the relative frequency approach. In the Bayesian approach probability is a measure of the ideal degree of confidence that an ideal perfectly rational mind would have about the state (or outcome) of the system (or random process) given only all the objectively available data extant for that system/process upon which to assign such a confidence measure on the various possible outcomes. The Bayesian approach is about the procedure of most rationally assigning a measure of the various degrees of confidence to the possible outcomes of some random process armed with just all the objectively available data.

> >In doing so, all of the strained attempts at creating a fictitious  
> >population out of measurements vanish (such as, say, analyzing  
> >measurements of the mass of the moon by imagining many hypothetical  
> >universes of "identical" measurements). On instead is quantifying  
> >your state of knowledge.

Again, it's not the subjective state of the actual knowledge of an actual less-than-completely-rational mind that is relevant. Rather it would be better considered to be the state of knowledge of an 'ideal' perfectly rational mind that is supplied with *only* *all* the objectively available data of the situation about which there is only partial information extant.

> >In almost all easy cases, the Bayesian approach yields the *exact*  
> >*same* numerical result as the standard approach. The interpretation  
> >is a lot easier, and a lot easier to communicate to students.  
> >  
> >

> > Brian Blais

David Bowman

-----  
 Forum for Physics Educators  
 Phys-1@carnot.physics.buffalo.edu  
<https://carnot.physics.buffalo.edu/mailman/listinfo/phys-1>

## 1.1 History

For a much more detailed account, please see (Loredo, 1990; Jaynes, 2003)

First formal account of the calculation of probabilities from Bernoulli (Bernoulli, 1713), who defined probability as a “degree of certainty”. His theorem states that, if the probability of an event is  $p$  then the limiting frequency of that event converges to  $p$ . It was later, by Bayes and Laplace that the inverse problem was solved: given  $n$  occurrences out of  $N$  trials, what is the probability  $p$  of a single occurrence?

The solution was published posthumously by Rev. Thomas Bayes (1763), and soon rediscovered, generalized, and applied to astrophysics by Laplace. It is Laplace who really brought probability theory to a mature state, applying it to problems in astrophysics, geology, meteorology, and others. One famous application was the determination of the masses of Jupiter and Saturn and the quantification of the uncertainties.

### 1.1.1 Derivation of Bayes’ Theorem

Laplace took as axioms the sum and product rules for probability:

$$\begin{aligned} p(A|C) + p(\bar{A}|C) &= 1 \\ p(AB|C) &= p(A|BC)p(B|C) \end{aligned}$$

from there, given the obvious symmetry  $p(AB|C) = p(BA|C)$  we get

$$\begin{aligned} p(A|BC)p(B) &= p(B|AC)p(A) \\ p(A|BC) &= \frac{p(B|AC)p(A)}{p(B)} \end{aligned}$$

which is Bayes’ Theorem.

### 1.1.2 Further History

After Laplace’s death, his ideas came under attack by mathematicians. They criticized two aspects of the theory:

1. The axioms, although *reasonable*, were not clearly unique for a definition of probability as vague as “degrees of plausibility”. The definition seemed vague, and thus the axioms which support the theory seemed arbitrary.

If one defines probabilities as limiting frequencies of events, then these axioms are justified.

2. It was unclear how to assign the prior probabilities in the first place. Bernoulli introduced the *Principle of Insufficient Reason*, which states that if the evidence does not provide any reason to choose  $A_1$  or  $A_2$ , then one assigns equal probability to both. If there are  $N$  such propositions, then the probability is assigned  $1/N$  for each. Although again, reasonable, it was not clear how to generalize it to a continuous variable.

If one defines probabilities as limiting frequencies of events, this problem disappears, because the notion of prior probabilities disappeared, as well as the probability of an hypothesis. Hypotheses are true or false (1 or 0) for *all* elements of an ensemble or repeated experiment, and thus does not have a limiting frequency other than 0 or 1.

### 1.1.3 Response

Shifting to a limiting frequency definition, researchers avoided the issues above, and did not pursue their direct solution vigorously. The solutions did come, however.

1. In the mid-1900’s, R. T. Cox (1946, 1961) and E. T. Jaynes (1957, 1958) demonstrated that, from a small collection of reasonable “desiderata” (aka axioms), one could develop a complete and rigorous mathematical theory from “degrees of plausibility”. These “desiderata” are:
  - (a) Degrees of plausibility are represented by real numbers
  - (b) Qualitative correspondence with common sense. Consistent with deductive logic in the limit of true and false propositions.
  - (c) Consistency
    - i. If a conclusion can be reasoned out in more than one way, every possible way must lead to the same result
    - ii. The theory must use all of the information provided
    - iii. Equivalent states of knowledge must be represented by equivalent plausibility assignments

With just these it is shown that the original, Laplace, methods of using Bayes’ theorem were the correct ones. It is also shown that **any theory of probability is either Bayesian, or violates one of the above desiderata.**

2. The concern about assigning prior probabilities was answered in the work of Shannon and Jaynes, with the advent of maximum entropy methods and the methods of transformation groups.

As a note, it is worth quoting (Loredo, 1990): “It is worth emphasizing that probabilities are *assigned*, not *measured*. This is because probabilities are measures of plausibilities of propositions;

they thus reflect whatever information one may have bearing on the truth of propositions, and are not properties of the propositions themselves.

...

“In this sense, Bayesian Probability Theory is ‘subjective,’ it describes states of knowledge, not states of nature. But it is ‘objective’ in that we insist that equivalent states of knowledge be represented by equal probabilities, and that problems be well-posed: enough information must be provided to allow unique, unambiguous probability assignments.”

Although there isn’t a unique solution for converting verbal descriptions into prior probabilities in *every* case, the current methods allow this translation in many very useful cases.

## 1.2 Procedure

In all of the cases that follow, there is a common procedure. We want to estimate parameters in a model, so we write down a probability distribution for those parameters dependent on the data, and any available information,  $I$ . If we have one parameter in the model, then the form is like:

$$p(\text{parameter}|\text{data}, I)$$

We apply Bayes’ theorem/rule to write it in terms of things we have a handle on:

$$p(\text{parameter}|\text{data}, I) = \frac{p(\text{data}|\text{parameter}, I)p(\text{parameter}|I)}{p(\text{data}|I)}$$

The left-hand side and the top two terms have names, and the bottom term is a normalization constant (which we will often omit, and work in proportions).

$$\begin{aligned} \underbrace{p(\text{parameter}|\text{data}, I)}_{\text{posterior}} &= \frac{\overbrace{p(\text{data}|\text{parameter}, I)}^{\text{likelihood}} \overbrace{p(\text{parameter}|I)}^{\text{prior}}}{\underbrace{p(\text{data}|I)}_{\text{normalization}}} \\ &\propto \underbrace{p(\text{data}|\text{parameter}, I)}_{\text{likelihood}} \overbrace{p(\text{parameter}|I)}^{\text{prior}} \end{aligned}$$

The likelihood is how the data could be generated from the model. The prior is a weighting of the parameter possibilities, before we see the data.

All of the information in the problem is complete once the posterior is written down. After that, it is a matter of working with that distribution to obtain the estimate. Often, we take the maximum posterior, but we can also take the mean, median or any other central measure. We can look at standard deviations to determine confidence intervals, but we can also look at quartiles. We will often look at the log of the posterior, as an analytical trick. When all else fails, we can find the estimate numerically from the posterior.

### 1.2.1 Orthodox Hypothesis Tests

Much of orthodox hypothesis testing can be interpreted in a much more straightforward way with the Bayesian approach. For example, the  $p$  value calculated in the orthodox fashion is “the probability, computed assuming that the null hypothesis  $H_o$  is true, of observing a value of the test statistic that is at least as extreme as the value actually computed from the data” (Bowerman and O’Connell, 2003). In the orthodox method, if you want to infer from the data that the mean value is, say, greater than zero you set up the null with  $H_o : \mu \leq 0$  and the alternate with  $H_a : \mu > 0$ , select the appropriate statistic ( $z$ ,  $t$ , etc. . . ), calculate the  $p$ -value of the null, where you use hypothetical data and look for the frequency that  $H_o$  is true. Finally, you reject the null at the level of significance, usually at the 5% level. No wonder students get confused!

In the Bayesian way, we take the posterior distribution for the parameter,  $\mu$ , and then ask “what is the probability that  $\mu$  is greater than 0?”, integrate the posterior probability distribution from 0 to infinity and get that probability directly. In many applications, *they both give identical results!*

(Jeffreys, 1939) put it in the following way:

What the use of  $p$  implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure. On the face of it the fact that such results have not occurred might be more reasonably be taken as evidence for the law, not against it.

Another comment, this time from Jaynes (I need to find the cite) is that, if we reject  $H_o$ , then we should reject probabilities conditional on  $H_o$  being true. The  $p$ -value is such a probability, so the procedure invalidates itself. Once I find the cite, I will replace with with Jaynes’ more eloquent description.

## 1.3 Numerical Analysis

Appendix D includes the actual data values for the different cases. In this section I include Python code for doing the numerical analysis on data from the textbook.

### 1.3.1 Plotting Distributions and Histograms

```
from utils import *

mu=10
sd=2
x=sd*randn(1000,1)+mu # make random numbers

figure(1)
clf()
plot(x,'o')
xlabel('data point')
ylabel('value')
```

```
title('artificial data')
savefig('art1_1.pdf')

n,v=hist(x,50) # 50 bins
figure(2)
clf()
bar(v,n,width=.3)
ylabel('number of points')
xlabel('value')
title('artificial data')
savefig('art1_2.pdf')

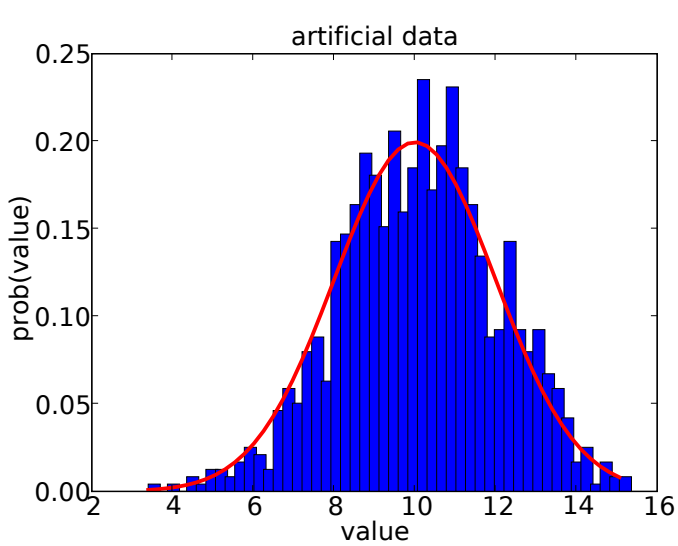
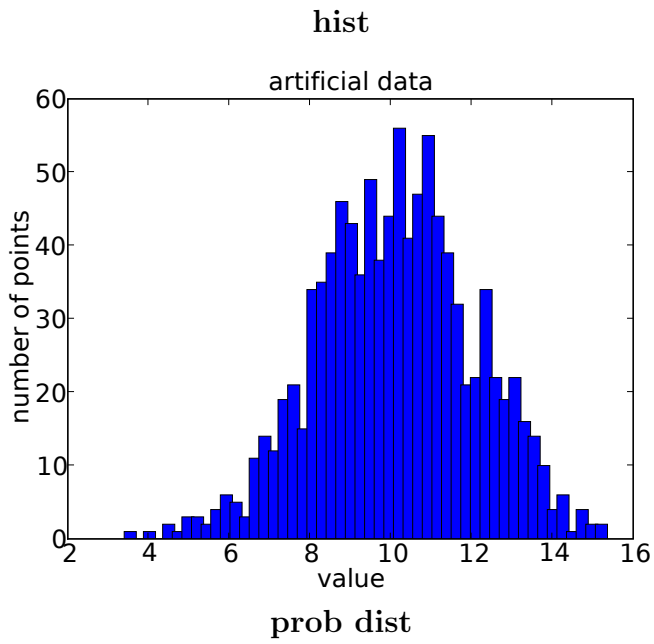
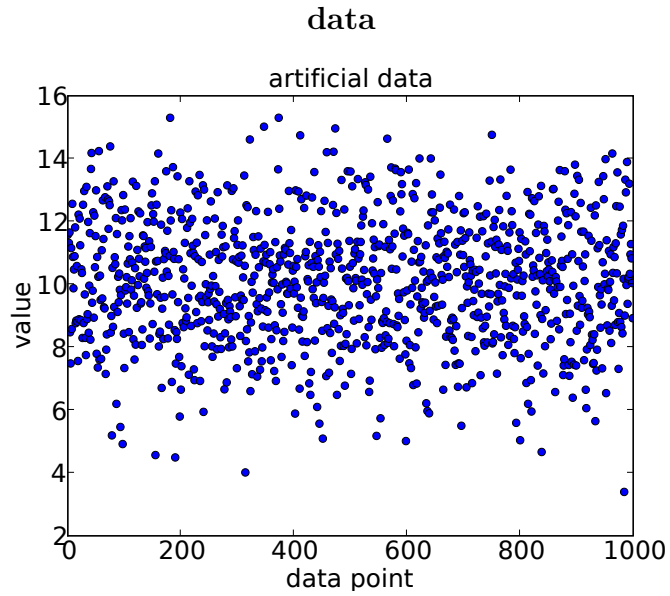
figure(3)
clf()

pv=1/sqrt(2*pi*sd**2)*exp(-(v-mu)**2/(2*sd**2))
plot(v,pv,'r-',linewidth=3)

dv=v[1]-v[0]
pv=n/dv/n.sum() # make probability distribution
bar(v,pv,width=.3)

ylabel('prob(value)')
xlabel('value')
title('artificial data')
savefig('art1_3.pdf')

show()
```



# Chapter 2

## One Sample Inferences

This chapter deals with the topics in the orthodox textbook in the following chapters: Chapter 6 (Sampling Distributions), Chapter 7 (Confidence Intervals) and Chapter 8 (Hypothesis Testing).

### 2.1 Unknown $\mu$ , Known $\sigma$

My first difficulty with the orthodox textbook is in Chapter 6: “Sampling Distributions”. I looked at the first section, “The Sampling Distribution of the Sample Mean”, and had to re-read it several times before I understood what was going on. I personally find that the Bayesian approach is simpler and more intuitive.

The orthodox approach defines probability as the long-term frequency of events. Because of this definition, it makes no sense to talk about the probability of a parameter (because the parameter is supposedly fixed). Instead one talks about sampling distributions of estimates of the parameter.

In the Bayesian approach, one can ask about the probability of a parameter, because probability is defined as the real-number plausibility value, as an extension of deductive logic. Thus one can ask, in my opinion, the simpler question: what is the probability of the mean given the data ( $p(\mu|D)$ ).

From Bayes’ theorem, we can write down the prior, the likelihood, do some math, and obtain the posterior. The full derivation is given in Appendix B.1.

#### (Uniform) Prior

$$p(\mu|\sigma, I) = p(\mu|I) = \begin{cases} A & \mu_{\min} \leq \mu \leq \mu_{\max} \\ 0 & \text{otherwise} \end{cases}$$

#### Likelihood

$$p(\mathbf{x}|\mu, \sigma, I) = \prod_{k=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_k - \mu)^2 / 2\sigma^2}$$

## Posterior

$$p(\mu|\mathbf{x}, \sigma, I) = \sqrt{\frac{N}{2\pi\sigma^2}} e^{-N(\bar{x}-\mu)^2/2\sigma^2} \quad (2.1.1)$$

## Maximum Posterior Estimate

$$\begin{aligned} \mu &= \frac{\sum_{k=1}^N x_k}{N} \pm \frac{\sigma}{\sqrt{N}} \\ &\equiv \bar{x} \pm \frac{\sigma}{\sqrt{N}} \end{aligned} \quad (2.1.2)$$

This is the same result quoted in the orthodox textbook for unknown  $\mu$  but known  $\sigma$ . In this case, however, all of the information pertinent to the problem is in the posterior distribution. We can find the mean, median, mode, etc. To perform a hypothesis test, we simply do straightforward integrals on the posterior:

Test	Question	Value
one-tailed	What is the probability that $\mu > a$ ?	$p_H = \int_a^\infty p(\mu D, \sigma, I)d\mu$
one-tailed	What is the probability that $\mu < a$ ?	$p_H = \int_{-\infty}^a p(\mu D, \sigma, I)d\mu$
two-tailed <sup>1</sup>	What is the probability that $\mu \neq a$ ?	$d \equiv \text{abs}(\bar{x} - a)$ , $p_H = 1 - \int_{\bar{x}-d}^{\bar{x}+d} p(\mu D, \sigma, I)d\mu$

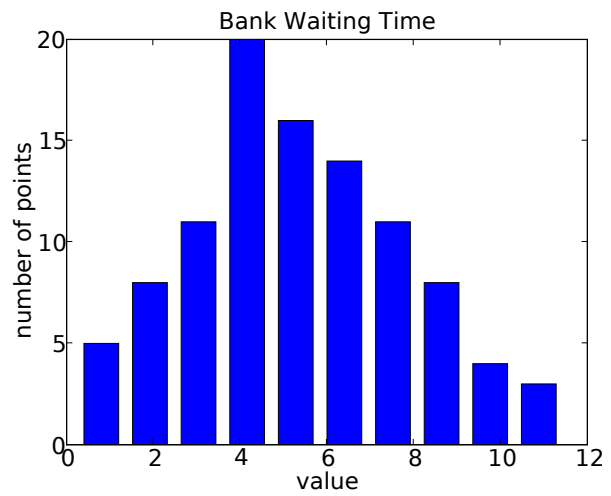
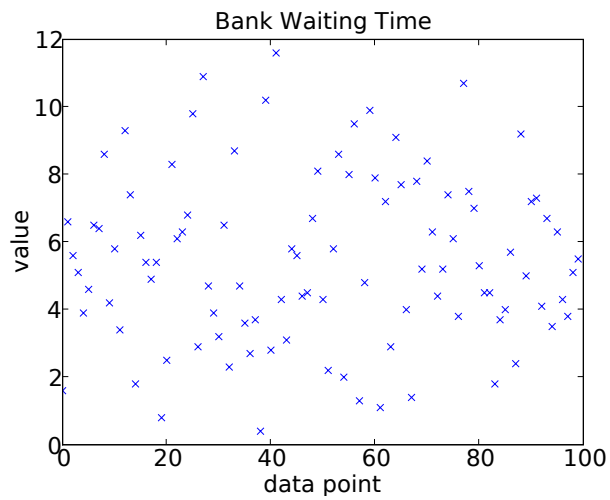
The examples which follow are also done in the textbook, with the same answers, but different interpretation.

### 2.1.1 Exercises

#### 2.1.1.1 Exercise 7.7: Bank Wait Times

Mean, Std, 95%, 99% confidence intervals. Less than 6?

### Preliminaries



```

from utils import *

y=get_data('WaitTime')

figure(1)
clf()
plot(y,'x')
xlabel('data point')
ylabel('value')
title('Bank Waiting Time')

savefig('fig071705_2a.pdf')

figure(2)
n,v=hist(y)
clf()
bar(v,n)

hist(y)
ylabel('number of points')
xlabel('value')
title('Bank Waiting Time')

savefig('fig071705_2b.pdf')
show()

```

### Output:

```
y_bar = 5.46
```

```
s = 0.246304689359
95.0% interval: [4.9773,5.9427]
99.0% interval: [4.8256,6.0944]
Probability of mu>=6: 1.42%
Probability of mu<=6: 98.58%
Number of standard deviations away: 2.19
```

**Code:**

```
from utils import *
from scipy import stats

# get the data
y=get_data('WaitTime')

N=len(y)

# best estimate for the value
#
y_bar=mean(y)
s=std(y)/sqrt(N)
print "y_bar =",y_bar
print "s =",s
# confidence levels
#
c=0.95
num_std=abs(stats.norm.ppf((1-c)/2,0,1))
pe=(stats.norm.cdf(num_std)-stats.norm.cdf(-num_std))*100

print "%.1f%% interval: [%.4f,%.4f]" % (c*100,y_bar-num_std*s,y_bar+num_std*s)

c=0.99
num_std=abs(stats.norm.ppf((1-c)/2,0,1))
pe=(stats.norm.cdf(num_std)-stats.norm.cdf(-num_std))*100

print "%.1f%% interval: [%.4f,%.4f]" % (c*100,y_bar-num_std*s,y_bar+num_std*s)

# probability of mu >=6
#
num_std=(6-y_bar)/s
print 'Probability of mu>=6: %.2f%%' % ((1-stats.norm.cdf(num_std))*100)
print 'Probability of mu<=6: %.2f%%' % (stats.norm.cdf(num_std)*100)
print 'Number of standard deviations away: %.2f' % abs(num_std)
```

### 2.1.1.2 Exercise 7.8: Customer Satisfaction

Mean, Std, 95%, 99% confidence intervals. Greater than 42?

#### Preliminaries



```

from utils import *

y=get_data('CustSat')

figure(1)
clf()
plot(y,'x')
xlabel('data point')
ylabel('value')
title('Customer Satisfaction Ratings')

savefig('fig071705_4a.pdf')

figure(2)
n,v=hist(y)
clf()
bar(v,n)

hist(y)
ylabel('number of points')
xlabel('value')
title('Customer Satisfaction Ratings')

savefig('fig071705_4b.pdf')

```

```
show()
```

**Output:**

```
y_bar = 42.9538461538
s = 0.325216919311
95.0% interval: [42.3164,43.5913]
99.0% interval: [42.1161,43.7915]
Probability of mu>=42: 99.83%
Probability of mu<=42: 0.17%
Number of standard deviations away: 2.93
```

**Code:**

```
from utils import *
from scipy import stats

# get the data
y=get_data('CustSat')

N=len(y)

# best estimate for the value
#
y_bar=mean(y)
s=std(y)/sqrt(N)
print "y_bar =",y_bar
print "s =",s
# confidence levels
#
c=0.95
num_std=abs(stats.norm.ppf((1-c)/2,0,1))
pe=(stats.norm.cdf(num_std)-stats.norm.cdf(-num_std))*100

print "%.1f%% interval: [%.4f,%.4f]" % (c*100,y_bar-num_std*s,y_bar+num_std*s)

c=0.99
num_std=abs(stats.norm.ppf((1-c)/2,0,1))
pe=(stats.norm.cdf(num_std)-stats.norm.cdf(-num_std))*100

print "%.1f%% interval: [%.4f,%.4f]" % (c*100,y_bar-num_std*s,y_bar+num_std*s)

# probability of mu >=42
#
```

```

num_std=(42-y_bar)/s
print 'Probability of mu>=42: %.2f%%' % ((1-stats.norm.cdf(num_std))*100)
print 'Probability of mu<=42: %.2f%%' % (stats.norm.cdf(num_std)*100)
print 'Number of standard deviations away: %.2f' % abs(num_std)

```

## 2.2 Unknown $\mu$ , Unknown $\sigma$

### (Uniform) Prior

$$p(\mu, \sigma | I) = \begin{cases} \text{constant} & \sigma > 0 \\ 0 & \text{otherwise} \end{cases}$$

### Likelihood

$$p(\mathbf{x} | \mu, \sigma, I) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^N e^{-\frac{1}{2\sigma^2} \sum_{k=1}^N (x_k - \mu)^2}$$

### Posterior

$$p(\mu, \sigma | \mathbf{x}, I) \propto \begin{cases} \left( \frac{1}{\sigma} \right)^N e^{-\frac{1}{2\sigma^2} \sum_{k=1}^N (x_k - \mu)^2} & \sigma > 0 \\ 0 & \text{otherwise} \end{cases}$$

### Posterior for $\mu$

$$p(\mu | \mathbf{x}, I) = \int_0^\infty p(\mu, \sigma | \mathbf{x}, I) d\sigma \quad (2.2.3)$$

$$\propto \left[ \sum_{k=1}^N (x_k - \mu)^2 \right]^{-(N-1)/2} \quad (2.2.4)$$

or, rewritten,

$$\sum_{k=1}^N (x_k - \mu)^2 \equiv N(\bar{x} - \mu)^2 + V$$

$$V \equiv \sum_{k=1}^N (x_k - \bar{x})^2$$

$$p(\mu | \mathbf{x}, I) \propto \left[ N(\bar{x} - \mu)^2 + V \right]^{-(N-1)/2} \quad (2.2.5)$$

which is the Student-t distribution with  $N - 2$  degrees of freedom. Had we used the suggested (Jeffreys) prior for the scale parameter,  $\sigma$ , we would get

$$p(\mu|\mathbf{x}, I) \propto [N(\bar{x} - \mu)^2 + V]^{-N/2} \quad (2.2.6)$$

which is the Student-t distribution with  $N - 1$  degrees of freedom.

We now define

$$t \equiv \frac{\mu - \bar{x}}{S/\sqrt{N}}$$

This variable  $t$  has exactly the standard t-distribution with  $N - 1$  degrees of freedom, in a form which can be readily looked up.

### Maximum Posterior Estimate for $\mu$ (Gaussian Confidence)

$$\begin{aligned} \mu &= \bar{x} \pm \frac{S}{\sqrt{N}} \\ \bar{x} &\equiv \frac{1}{N} \sum_{k=1}^N x_k \\ S^2 &\equiv \frac{1}{(N-1)} \sum_{k=1}^N (x_k - \bar{x})^2 \end{aligned}$$

### Posterior for $\sigma$

$$p(\sigma|\mathbf{x}, I) = \int_{-\infty}^{\infty} p(\mu, \sigma|\mathbf{x}, I) d\mu \quad (2.2.7)$$

$$\propto \frac{1}{\sigma^{N-1}} e^{-V/2\sigma^2} \quad (2.2.8)$$

for the Jeffreys' prior, this becomes

$$p(\sigma|\mathbf{x}, I) \propto \frac{1}{\sigma^N} e^{-V/2\sigma^2} \quad (2.2.9)$$

or, changing variables to  $\xi \equiv V/\sigma^2$ , then this is the  $\chi^2$  distribution with  $f \equiv N - 1$  degrees of freedom:

$$\begin{aligned} p(\xi|\mathbf{x}, I) &\propto \xi^{\frac{N-1}{2}-1} e^{-\xi/2} \\ &\propto \xi^{\frac{f}{2}-1} e^{-\xi/2} \end{aligned}$$

### Maximum Posterior Estimate for $\sigma$ (Gaussian Confidence)

$$\sigma = S^2 \pm \frac{S^2}{\sqrt{2(N-1)}} \quad (2.2.10)$$

$$S^2 \equiv \frac{1}{(N-1)} \sum_{k=1}^N (x_k - \bar{x})^2 \quad (2.2.11)$$

## 2.2.1 Exercises

### 2.2.1.1 Exercise 7.24: Bank Wait Times

Using t-based confidence intervals around the mean, and median confidence intervals. In the Bayesian method, these seem to be the same.

#### Output:

```
mean: 5.4600
95.0% t-dist interval: [4.9688,5.9512]
99.0% t-dist interval: [4.8098,6.1102]
median: 5.4600
95.0% t-dist median interval: [4.9690,5.9510]
```

#### Code:

```
from utils import *
from scipy import stats

# get the data
y=get_data('WaitTime')

N=len(y)

# best estimate for the value
#
y_bar=mean(y)
V=sum((y-y_bar)**2)
s=sqrt(V/(N-1))
print "mean = %.4f" % y_bar

# confidence levels
#
```

```

c=0.95
bottom=stats.t.ppf((1-c)/2,N-1)
top=stats.t.ppf(c+(1-c)/2,N-1)
pe=(stats.t.cdf(top,N-1)-stats.t.cdf(bottom,N-1))*100

print "%.1f%% t-dist interval: [%.4f,%.4f]" % (pe,
                                             y_bar+bottom*s/sqrt(N),
                                             y_bar+top*s/sqrt(N))

# confidence levels
#
c=0.99
bottom=stats.t.ppf((1-c)/2,N-1)
top=stats.t.ppf(c+(1-c)/2,N-1)
pe=(stats.t.cdf(top,N-1)-stats.t.cdf(bottom,N-1))*100

print "%.1f%% t-dist interval: [%.4f,%.4f]" % (pe,
                                             y_bar+bottom*s/sqrt(N),
                                             y_bar+top*s/sqrt(N))

# work with the pdf directly for the median
mu=r_[0:20:.001]
dmu=mu[1]-mu[0]

p_mu=(N*(y_bar-mu)**2+V)**(-N/2.0)
p_mu=p_mu/sum(p_mu)/dmu
cp_mu=cumsum(p_mu*dmu)

idx=find(cp_mu>=.5)
print "median: %.4f" % mu[idx[0]]

c=.95

idx=find(cp_mu>=(1-c)/2)
bottom=mu[idx[0]]
idx=find(cp_mu>=((1-c)/2+c))
top=mu[idx[0]]

print "%.1f%% t-dist median interval: [%.4f,%.4f]" % (c*100,bottom,top)

```

## 2.3 Unknown proportion

This analysis is good for situations like determining if a coin is fair. We are trying to estimate a proportion,  $\theta$ . For a coin,  $\theta = 0$  would be a coin that always falls heads,  $\theta = 1$  would be a coin that always falls tails, and  $\theta = 1/2$  would be a fair coin.

If  $\theta$  is the model representing the probability,  $\theta$ , of the coin landing on heads (and  $1 - \theta$  is the probability of landing on tails), we need to make an estimate of probability of model  $\theta$  being true given the data, which will consist of  $N$  flips of which  $h$  are heads. The full derivation is done in Appendix B.3.

### (Uniform) Prior

$$p(\theta|I) = 1$$

### Likelihood

$$p(D|\theta) = \binom{N}{h} \theta^h (1 - \theta)^{N-h}$$

### Posterior

$$p(\theta|D, I) = (N + 1) \cdot \binom{N}{h} \theta^h (1 - \theta)^{N-h} \quad (2.3.12)$$

$$= \frac{(N + 1)!}{h!(N - h)!} \theta^h (1 - \theta)^{N-h} \quad (2.3.13)$$

### Maximum Posterior Estimate

$$\theta = \frac{h}{N} \quad (2.3.14)$$

### Mean, Standard Deviation Estimate

$$\bar{\theta} = \frac{h + 1}{N + 2} \quad (2.3.15)$$

$$\sigma^2 = \bar{\theta}(1 - \bar{\theta}) \frac{1}{N + 3} \quad (2.3.16)$$

If we define  $f \equiv h/N$ , and take the limit for large  $N$  and  $fN$ , then we have the approximation

$$\begin{aligned}\bar{\theta} &\approx \frac{h}{N} \\ \sigma^2 &\approx \frac{f(1-f)}{N} \\ p(\theta|D, I) &\approx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\theta-\bar{\theta})^2/2\sigma^2}\end{aligned}$$

### 2.3.1 Confidence

We want to know how many flips to do to have confidence in our estimate. What this means is that we want to know something like “I am 95% confident that the correct model falls between  $\theta = 0.39$  and  $\theta = 0.41$ .” The range I will call  $E$ , and the confidence level will be determined by a parameter called  $Z$  which determines how many standard deviations from the mean am I including in my range.  $Z = 2$  is approximately 95% confidence.<sup>2</sup>

We now have

$$\begin{aligned}Z\sigma &= E \\ \sigma^2 &\approx \frac{f(1-f)}{N} \\ \frac{E^2}{Z^2} &\approx \frac{f(1-f)}{N} \\ N &\approx f(1-f) \left( \frac{Z^2}{E^2} \right)\end{aligned}$$

Now, if instead of an absolute error range, we wanted to look at a *percent* error compared with  $f$ , then the range we would look at would be  $Ef$  where  $E$  is now a percent difference. The number of flips necessary would now be

$$\begin{aligned}Z\sigma &= fE \\ \sigma^2 &\approx \frac{f(1-f)}{N} \\ \frac{E^2 f^2}{Z^2} &\approx \frac{f(1-f)}{N} \\ N &\approx (1-f) \left( \frac{Z^2}{fE^2} \right)\end{aligned}$$

The first relationship has a peak at  $f = 0.5$ , which is somewhat counter-intuitive. The second relationship diverges at  $f = 0$ , meaning that you need more flips to have the same percentage confidence in lower probability events.

---

<sup>2</sup>Technically,  $Z = \text{erfinv}(\text{percentile}) \cdot \sqrt{2}$  for the Gaussian distribution.

If you are interested in the confidence range one has, given a certain number of flips, then we have

$$E \approx Z \sqrt{\frac{(1-f)}{fN}}$$

where your range is  $f \pm E$ . Again,  $Z = 2$  is about 95 percent confidence.

Why is this not symmetric for 0 and 1? If you have 10 percent error on a probability estimate of 0.9, you could be off by 0.09. This amount of deviation would correspond to 90 percent error on the remaining 0.1.

### 2.3.2 Median and Percentiles

Because the beta distribution is non-symmetric, especially at the probability extremes (around 0 or 1), then the mean may not be the best estimate to use for the probability. A better estimate in these cases is the median of the distribution. Unfortunately there isn't an easy analytical solution to the problem of finding the median,  $\hat{\theta}$ , in the following:

$$\begin{aligned} p(\theta|D, I) &= \frac{(N+1)!}{h!(N-h)!} \theta^h (1-\theta)^{N-h} \\ \frac{1}{2} &= \int_0^{\hat{\theta}} \frac{(N+1)!}{h!(N-h)!} \theta^h (1-\theta)^{N-h} d\theta \end{aligned}$$

Numerically, however, this is quite easy. The following Python code returns the proper value for  $\hat{\theta}$  that below which the probability distribution contains an area of  $A$ . For the median,  $A = 0.5$ .

```
from scipy.stats import beta
```

```
def percentile_beta(h,N,p):
```

```
    a=h+1
```

```
    b=(N-h)+1
```

```
    y=beta.ppf(p,a,b)
```

```
    return y
```

To get the median, we use

```
median=percentile_beta(h,N,.5)
```

For example,

```
In [1]:percentile_beta(5,10,0.5)
```

```
Out[1]:array(0.5)
```

```
In [2]:percentile_beta(1,10,0.5)
```

```
Out[2]:array(0.14796342543060625)
```

To get a confidence range of 68% (i.e. 68% of the distribution lies within the range, 34% above the median, and 34% below) we do the following:

```
In [3]:c=0.68 # confidence range
In [4]:m1=percentile_beta(h,N,.5-c/2) # lower bound
In [5]:m2=percentile_beta(h,N,.5+c/2) # upper bound
```

For example, with  $h = 2$ ,  $N = 10$  we have

- Median  $\pm 34\%$ :  $0.129 < 0.236 < 0.372$
- Mean  $\pm \sigma$ :  $0.074 < 0.2 < 0.326$

Empirically, it doesn't look like the median gets us very far in this case: it's almost the same as the mean, except for small probabilities.

### 2.3.3 Numerical Examples

#### 2.3.3.1 Is this a fair coin?

The probability of  $\theta$  being within 10% of the fair value of 0.5 (so, between 0.45 and 0.55) is

$$\int_{0.45}^{0.55} p(\theta|D, I) d\theta$$

which is found by looking at the beta distribution cdf.

```
In [1]:h=2; N=10
```

```
In [2]:a=h+1; b=(N-h)+1
```

```
In [3]:stats.beta.cdf(.55,a,b)-stats.beta.cdf(.45,a,b)
Out[3]:0.0504209391752
```

So there is approximately a 5% chance of it being fair, defined by a  $\pm 10\%$  range around fair.

#### 2.3.3.2 Simple vs Complex Hypothesis Tests

Although I don't have enough space to cover this in detail, I wanted to give an example of the following comparison:

- $H_0$ : fair coin ( $\theta = 0.5$ )
- $H_1$ : coin with probability,  $\theta$

Here we are comparing a model with zero parameters versus a model with an adjustable parameter. Maximum likelihood will always choose the one with more parameters, because it can always fit the data better. This is the “variance” part in the “bias-variance” problem, similar to the overfitting problem in regression. The Bayesian method is interesting because it includes an Ockham factor, penalizing more complex models *automatically*. For example, this is what we have for the coin problem.

The odds ratio, or ratio of posteriors, is

$$B = \frac{p(H_1|D)}{p(H_o|D)}$$

which is the ratio of the probability for the complex model to the simple one. If  $B \gg 1$ , then we strongly prefer the complex one. If  $B \ll 1$  then we prefer the simple one.

$$\begin{aligned} B &= \frac{p(H_1|DI)}{p(H_o|DI)} \\ &= \underbrace{\frac{p(D|H_1I)}{p(D|H_oI)}}_{\text{likelihood ratio}} \times \underbrace{\frac{p(H_1|I)}{p(H_o|I)}}_{\text{prior ratio}} \end{aligned}$$

Say we have no reason to prefer either model

$$p(H_o|I) = p(H_1|I) = \frac{1}{2}$$

For the simple, fair coin model we have

$$p(D|H_oI) = \binom{N}{h} \left(\frac{1}{2}\right)^N$$

For the complex model we have to integrate over the adjustable parameters

$$\begin{aligned} p(D|H_1I) &= \int_0^1 p(D|\theta, H_1, I) p(\theta|H_1, I) d\theta \\ &= \int_0^1 \binom{N}{h} \theta^h (1-\theta)^{N-h} d\theta \\ &= \binom{N}{h} \text{Beta}(h+1, N-h+1) \end{aligned}$$

So now we have

$$\begin{aligned} B &= \frac{p(H_1|DI)}{p(H_o|DI)} \\ &= \text{Beta}(h+1, N-h+1) 2^N \end{aligned}$$

So, for  $h = 2$  and  $N = 10$  we have

```
In [1]:from scipy.special import beta
```

```
In [2]:h=2; N=10
```

```
In [3]:beta(h+1,N-h+1)*2**N
```

```
Out[3]:2.06868686869
```

or 2:1 for the more complex model.

If  $h = 4$  and  $N = 10$  we have

```
In [4]:h=4
```

```
In [5]:beta(h+1,N-h+1)*2**N
```

```
Out[5]:0.44329004329
```

```
In [6]:B=beta(h+1,N-h+1)*2**N
```

```
In [7]:1/B
```

```
Out[7]:2.255859375
```

or 2.2:1 against the more complex model, even though from maximum likelihood we would favor the  $\theta = 0.4$  solution, with likelihoods

```
In [1]:from utils import *
```

```
In [2]:.4**4*.6**6*nchoosek(10,4)
```

```
Out[2]:0.25082265599999998
```

```
In [3]:.5**4*.5**6*nchoosek(10,4)
```

```
Out[3]:0.205078125
```

# Chapter 3

## Two Sample Inferences

In this chapter we deal with inferences from two data sets, denoted  $\mathbf{x} \equiv \{x_1, x_2, \dots, x_n\}$  and  $\mathbf{y} \equiv \{y_1, y_2, \dots, y_m\}$ . Almost exclusively we are looking at a difference of means.

### 3.1 Paired Data Difference of Means, $\delta_k \equiv x_k - y_k$

In the case of paired data, where each of the  $x_k$  are paired with their corresponding  $y_k$  ( $n = m$ ). You can treat the  $\delta_k$  as a single sample, for the case of known and unknown  $\sigma$ .

### 3.2 Difference of Means, $\delta \equiv \mu_x - \mu_y$ , known $\sigma_x$ and $\sigma_y$

With a change of variables from  $\mu_x$  and  $\mu_y$  to  $\delta$ , we obtain a Gaussian posterior, with the obvious mean and a modified variance.

#### Posterior

$$\mu_\delta \equiv \mu_x - \mu_y \quad (3.2.1)$$

$$\sigma_\delta \equiv \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m} \quad (3.2.2)$$

$$p(\delta | \mathbf{x}, \mathbf{y}, \sigma_x, \sigma_y, I) = \frac{1}{\sqrt{2\pi\sigma_\delta^2}} e^{-(\delta - \mu_\delta)^2 / 2\sigma_\delta^2} \quad (3.2.3)$$

### 3.3 Difference of Means, $\delta \equiv \mu_x - \mu_y$ , unknown $\sigma_x$ and $\sigma_y$

Making definitions as before for the  $t$  distribution for each variable

$$t_x \equiv \frac{\mu_x - \bar{x}}{S_x / \sqrt{n}}$$

$$\begin{aligned}
t_y &\equiv \frac{\mu_y - \bar{y}}{S_y/\sqrt{n}} \\
S_x^2 &\equiv \frac{1}{(n-1)} \sum_{k=1}^n (x_k - \mu_x)^2 \\
S_y^2 &\equiv \frac{1}{(m-1)} \sum_{k=1}^m (y_k - \mu_y)^2
\end{aligned}$$

From the addition of variables we get

$$\begin{aligned}
t &\equiv \frac{\delta - (\bar{x} - \bar{y})}{\sqrt{S_x^2/m + S_y^2/n}} \\
\tan \theta &\equiv \frac{S_x/\sqrt{n}}{S_y/\sqrt{m}}
\end{aligned}$$

$\tan \theta$  depends on the data, and  $t_x$ , and  $t_y$  are known, so the distribution for  $t$  should be known. It is named the Behren's distribution.

An example from (Lee, 1989).

$$m = 12, n = 7, \bar{x} = 120, \bar{y} = 101, S_x^2 = 5032/11, S_y^2 = 2552/6.$$

$$\text{Thus, } \tan \theta = ((457/12)/(425/7))^{(1/2)} = 0.8, \sqrt{S_x^2/m + S_y^2/n} = 9.9$$

(Lee, 1989) then uses Behrens-Fisher tables to determine that the 90% confidence interval for the difference is between 0 and 38. We can do it another way, which does not directly use the Behrens-Fisher distribution. The method is used in (Jaynes, 1976) for a different problem which we outline below.

In the case where we are estimating the mean, and with unknown standard deviation, we arrive at the t-distribution.

$$p(\mu|\mathbf{x}, I) \propto [(\bar{x} - \mu)^2 + V]^{-N/2}$$

If we have two samples, we determine the sample means ( $\bar{x}$  and  $\bar{y}$ ) and sample standard deviations  $S_x$  and  $S_y$ , then we have

$$\begin{aligned}
p(\mu_x, \mu_y|\mathbf{x}, \mathbf{y}, I) &= p(\mu_x|\mathbf{x}, I)p(\mu_y|\mathbf{y}, I) \\
&\propto [(\bar{x} - \mu_x)^2 + S_x^2]^{-n/2} [(\bar{y} - \mu_y)^2 + S_y^2]^{-m/2}
\end{aligned}$$

Transforming to  $\mu \equiv \mu_x - \mu_y$  and  $\mu_x$  (the Jacobian is 1), and integrating out the  $\mu_x$  we get

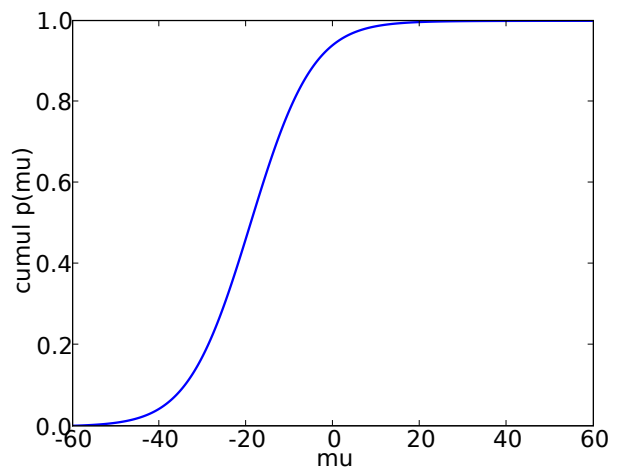
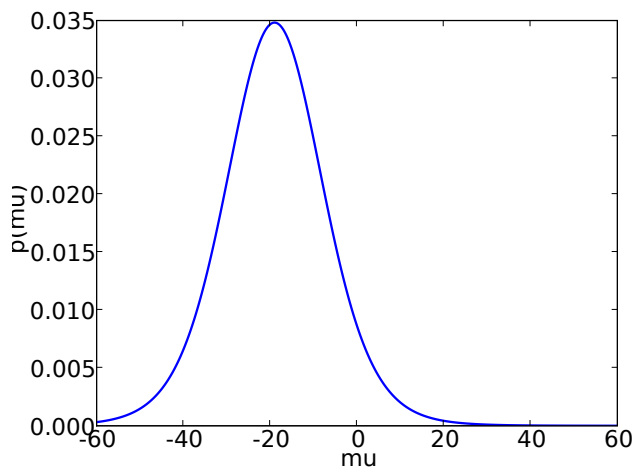
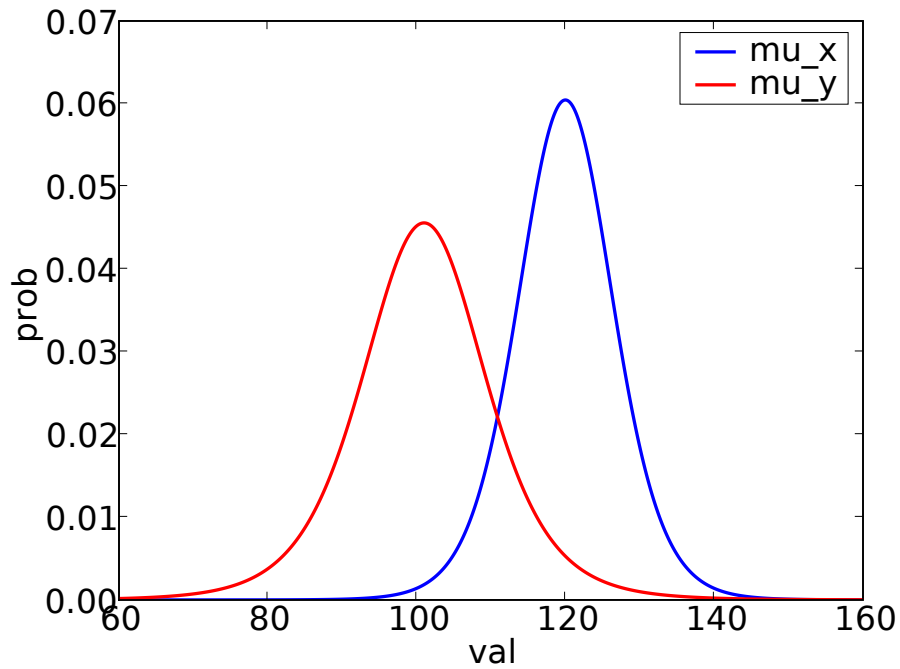
$$p(\mu|\mathbf{x}, \mathbf{y}, I) \propto \int_{-\infty}^{\infty} d\mu_x [(\bar{x} - \mu_x)^2 + S_x^2]^{-n/2} \times [(\bar{y} - (\mu_x - \mu))^2 + S_y^2]^{-m/2}$$

We can perform this integral numerically, by brute-force, quite simply. For the problem of (Lee, 1989) we have

**Output:**

Median: -18.9500

90% interval: [-38.9000:1.3000]

**Code:**

```
from utils import *  
  
# individual T-dists  
na=12
```

```
mu_a=120
s_a=21.4

nb=7
mu_b=101
s_b=20.6

da=.05
a=r_[60:160:da]
pa=(s_a**2+(a-mu_a)**2)**(-na/2.0)
norm_a=sum(pa)*da

db=.02
b=r_[60:160:db]

dmu=0.05
mu=r_[-60:60:dmu]
mu_orig=mu

pb=(s_b**2+(b-mu_b)**2)**(-nb/2.0)
norm_b=sum(pb)*db

pa=(s_a**2+(a-mu_a)**2)**(-na/2.0)/norm_a
pb=(s_b**2+(b-mu_b)**2)**(-nb/2.0)/norm_b
figure(1)
clf()
plot(a,pa,'b',b,pb,'r',linewidth=2)
ylabel('prob')
xlabel('val')
legend(['mu_x', 'mu_y'])
savefig('doit061605_1_1.pdf')

# mu

mu,a=meshgrid(mu,a)
pa=(s_a**2+(a-mu_a)**2)**(-na/2.0)/norm_a
pb=(s_b**2+(mu+a-mu_b)**2)**(-nb/2.0)/norm_b

p_mu=sum(pa*pb,axis=0)*da

mu=mu_orig
figure(2)
clf()
plot(mu,p_mu,linewidth=2)
ylabel('p(mu)')
```

```

xlabel('mu')
savefig('doit061605_1_2.pdf')

# cumulative

cp_mu=cumsum(p_mu*dmu)
figure(3)
plot(mu,cp_mu,linewidth=2)
ylabel('cumul p(mu)')
xlabel('mu')
savefig('doit061605_1_3.pdf')

idx=find(cp_mu>0.05); md_lower=mu[idx[0]]
idx=find(cp_mu>0.5); md=mu[idx[0]]
idx=find(cp_mu>0.95); md_upper=mu[idx[0]]

print 'Median: %.4f\n 90%% interval: [%.4f:%.4f]' % (md,md_lower,md_upper)

show()

```

### 3.3.1 Jaynes 1976 Difference of Means

There is a problem in (Jaynes, 1976) on the difference of means, handled in this way. The description of the problem is

“Two manufacturers, A and B, are suppliers for a certain component, and we want to choose the one which affords the longer mean life. Manufacturer A supplies 9 units for test, which turn out to have a (mean  $\pm$  standard deviation) lifetime of  $(42 \pm 7.48)$  hours. B supplies 4 units, which yield  $(50 \pm 6.48)$  hours.”

His notation for what was done above is particularly clear, if not completely explicit. In his example, he uses  $a$  and  $b$  for  $\mu_x$  and  $\mu_y$ . He also doesn't transform to the difference of means, but looks at the probability that  $b > a$ . This is written as

$$\text{Prob}(b > a) = \int_{-\infty}^{\infty} da \int_a^{\infty} db P_n(a) P_m(b)$$

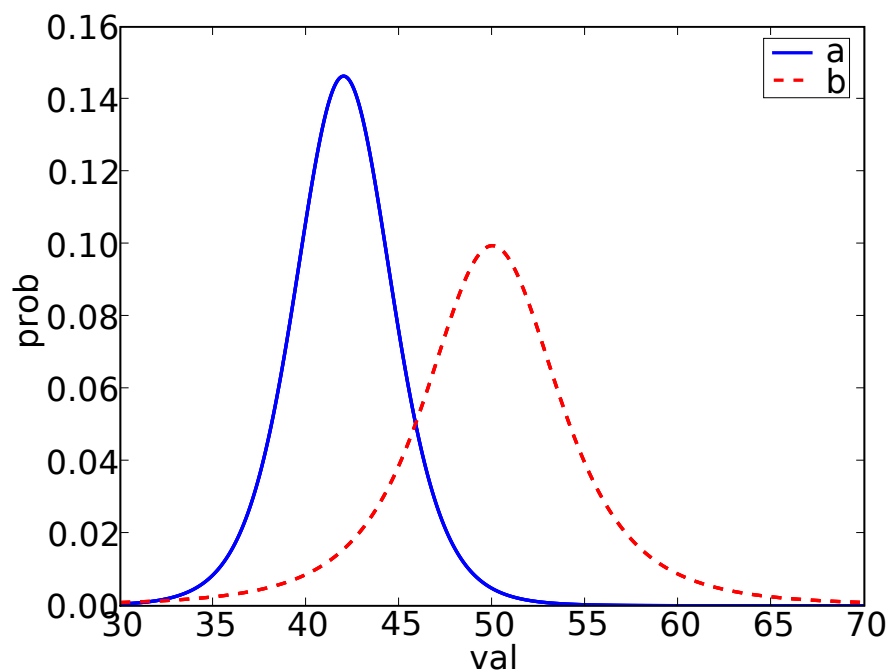
In our notation we have

$$\begin{aligned} \text{Prob}(\mu_y > \mu_x) &= \int_{-\infty}^{\infty} d\mu_x \int_{\mu_x}^{\infty} d\mu_y p(\mu_x | \mathbf{x}, I) p(\mu_y | \mathbf{y}, I) \\ &\propto \int_{-\infty}^{\infty} d\mu_x \int_{\mu_x}^{\infty} d\mu_y \left[ (\bar{x} - \mu_x)^2 + S_x \right]^{-n/2} \left[ (\bar{y} - \mu_y)^2 + S_y \right]^{-m/2} \end{aligned}$$

The conclusion is that we accept B being greater than A (at the 90% confidence level).

**Output:**

Probability for  $b > a = 0.9186$

**Code:**

```
from utils import *

na=9
mu_a=42
s_a=7.48

nb=4
mu_b=50
s_b=6.48

da=.05
a=r_[30:70:da]
pa=(s_a**2+(a-mu_a)**2)**(-na/2.0)
norm_a=sum(pa)*da

db=.05
b=r_[30:70:db]
pb=(s_b**2+(b-mu_b)**2)**(-nb/2.0)
norm_b=sum(pb)*db
```

```

pa=(s_a**2+(a-mu_a)**2)**(-na/2.0)/norm_a
pb=(s_b**2+(b-mu_b)**2)**(-nb/2.0)/norm_b

plot(a,pa,'b',b,pb,'r--',linewidth=2)
ylabel('prob')
xlabel('val')
legend(['a','b'])

a,b=meshgrid(a,b)
pa=(s_a**2+(a-mu_a)**2)**(-na/2.0)/norm_a
pb=(s_b**2+(b-mu_b)**2)**(-nb/2.0)/norm_b

idx=find(b.flat>a.flat)
p=sum(pa.flat[idx]*pb.flat[idx])*da*db

print 'Probability for b>a = %.4f' % p

savefig('fig061605_2.pdf')
show()

```

### 3.4 Ratio of Two Variances $\kappa \equiv \sigma_x^2/\sigma_y^2$

With the following definitions,

$$\begin{aligned}\eta &\equiv \kappa \times \frac{(V_y/f_y)}{(V_x/f_x)} \\ f_x &\equiv n - 1 \\ f_y &\equiv m - 1\end{aligned}$$

we get the posterior distribution,

**Posterior**

$$p(\eta|\mathbf{x}, \mathbf{y}, I) \propto \eta^{\frac{f_y}{2}-1} (f_x + f_y \eta)^{(f_x+f_y)/2} \quad (3.4.4)$$

which is the commonly used F distribution.

### 3.5 Simple Linear Regression, $y_k = mx_k + b + \epsilon$

Given

$$y_k = mx_k + b + \epsilon$$

where the (known) noise term is

$$p(\epsilon|I) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\epsilon/2\sigma^2}$$

then we have

### Posterior

$$p(m, b|\mathbf{y}, I) \propto \frac{1}{\sigma^N} e^{-\sum (mx_k + b - y_k)^2 / 2\sigma^2} \quad (3.5.5)$$

Maximizing with respect to  $m$  and  $b$  we get

$$\begin{aligned} m &= \frac{c - N\bar{x}\bar{y}}{v - N(\bar{x})^2} \\ b &= \frac{v\bar{y} - c\bar{x}}{v - N(\bar{x})^2} \end{aligned}$$

with

$$\begin{aligned} v &\equiv \sum x_k^2 \\ c &\equiv \sum x_k y_k \end{aligned}$$

## 3.6 Linear Regression with Errors on both $x$ and $y$

**note:** this is a sketch from (Jaynes, 1976). I haven't gone through this at all. I just wanted to have them as notes

The model is

$$Y_i = \alpha + \beta X_i$$

with measured values  $x_i = X_i + e_i$  and  $y_i = Y_i + f_i$ , with  $e_i \sim N(0, \sigma_x)$  and  $f_i \sim N(0, \sigma_y)$ .  $\sigma_x$  and  $\sigma_y$  unknown. Data is  $D = \dots(x_i, y_i)\dots$

Likelihood function

$$L(\alpha, \beta, \sigma_x, \sigma_y, X_i) = (\sigma_x \sigma_y)^{-n} e^{-\frac{1}{2} \sum_{i=1}^n \left( \frac{(x_i - X_i)^2}{\sigma_x^2} + \frac{(x_i - \alpha - \beta X_i)^2}{\sigma_y^2} \right)}$$

Need to integrate out nuisance parameters  $X_i$ ,  $\sigma_x$  and  $\sigma_y$ . Doing the  $X_i$  (with uniform prior) we get a function which only depends on

$$Q(\alpha, \beta) \equiv \frac{1}{n} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

Change variables from  $(\sigma_x, \sigma_y)$  to  $(\sigma, \lambda)$  where  $\sigma^2 = \sigma_y^2 + \beta^2 \sigma_x^2$  and  $\lambda = \sigma_y / \sigma_x$  (essentially noise perpendicular to the line and parallel to the line, respectively). Posterior on  $(\alpha, \beta)$  is found to be independent of  $\lambda$ , which is what we would expect (noise parallel to the line has no effect on the parameters of the line).

Integrating the  $\sigma$ , we are left with the function

$$f(\alpha, \beta) \sim Q(\alpha, \beta)^{-n/2}$$

which, when multiplied by the prior and normalized, gives the joint posterior.

### 3.7 Goodness of Fit

**(note: this derivation is still hazy to me, and seems incorrect. I'll look into it later) the improved goodness of fit is used down below.**

If we have a binomial process, where we observe  $n_1$  instances of a total of  $N$  observations, each instance having some probability  $p_1$ , then we have (with a uniform prior)

$$\begin{aligned} p(p_1 | n_1, I) &\propto p_1^{n_1} (1 - p_1)^{N - n_1} \\ \text{for large } N &\approx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(n_1 - Np_1)^2 / 2\sigma^2} \end{aligned}$$

where

$$\sigma^2 = Np_1(1 - p_1)$$

For  $k$  such possible instances, with occurrences  $n_1, n_2, \dots, n_k$  and probabilities  $p_1, \dots, p_k$  we have

$$p(\mathbf{p} | \mathbf{n}, I) \approx \prod \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-(n_i - Np_i)^2 / 2\sigma_i^2}$$

We have seen earlier that the variable  $\xi \equiv \sum (n_i - \bar{n})^2 / \sigma_i^2$  has a posterior  $\chi^2$  distribution. This sum can be rewritten as

$$\xi = \sum \frac{(n_i - Np_i)^2}{Np_i}$$

or, if the *expected* number is  $E_i = Np_i$ , then we have

$$\xi = \sum \frac{(n_i - E_i)^2}{E_i} \equiv \chi^2$$

Thus, in the large  $N$  limit, this value represents a “goodness of fit” measure.

An example is the following (from (Loredo, 1990)). Suppose there is a prediction that the fraction of Type A stars is  $a = 0.1$  in a cluster. An astronomer measures 5 Type A stars out of 96, and is wondering whether the prediction is supported.

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(5 - 9.6)^2}{9.6} + \frac{(91 - 86.4)^2}{86.4} = 2.45\end{aligned}$$

The probability that  $\chi^2$  is greater than 2.45 is

```
>> 1-chisquare_cdf(2.45,1)
ans = 0.11752
```

So the prediction is acceptable (was not rejected) at the 95% confidence level.

### 3.7.1 Jaynes' Alternative to $\chi^2$

In (Jaynes, 2003) an alternative to the  $\chi^2$  goodness of fit is proposed. He notes that if  $x \equiv n_k/Np_k$  and  $\log x \geq (1 - x^{-1})$  with equality if and only if  $x = 1$ , we get

$$\sum n_k \log \left( \frac{n_k}{Np_k} \right) \geq 0$$

Thus, he is motivated to suggest the following goodness of fit (working in base 10 “decibels”)

$$\psi \equiv 10 \sum n_k \log_{10} \left( \frac{n_k}{Np_k} \right)$$

Note that if the hypothesis is true, and the observed  $n_k$  is close to the expected  $Np_k$  then we have

$$\sum n_k \log \left( \frac{n_k}{Np_k} \right) = \frac{1}{2} \sum \frac{(n_k - Np_k)^2}{Np_k} + O\left(\frac{1}{\sqrt{n}}\right)$$

or, the  $\chi^2$  measure of goodness-of-fit. In this approximation,  $\psi \approx 10 \log_{10}(e) \times \frac{1}{2} \chi^2 \approx 4.343 \chi^2$

Each 10 decibels increase is a factor of 10 odds against, so it is relatively easy to compare different hypotheses.

# Chapter 4

## Orthodox versus Bayesian Approaches

### 4.1 Flipping a Tack

The idea, and data, for this problem comes from (Lindley and Phillips, 1976). The problem is very simple, and as such is very educational. The experimenter took a thumb-tack, flipped it onto a table, and noted whether the point was up or down against the table. He obtained the following data from repeated flips:

UUUDUDUUUUUD - (9 Ups, and 3 Downs)

The experimenter then wants to “assess the chance that the tack will fall ‘Up’ on a further, thirteenth, similar toss”. Or, perhaps a more easily answered question, “is there good evidence that this tack is (or is not) unbiased (50-50 chance of U or D)?”

#### 4.1.1 Orthodox Statistics

You would think that this problem would have a unique and straightforward solution using orthodox statistics. After all, it is about the simplest problem for which one can apply statistics, and one of the first problems toward which probability theory was brought to bear. In reality, without further information of a dubious nature, there is no unique orthodox solution to this problem. A standard approach would work something like the following:

We set up a hypothesis that the coin is unbiased. One obtains a p-value which gives “the chance of the observed result or more extreme”. Results that are more extreme would be

- 10 U + 2 D
- 11 U + 1 D
- 12 U + 0 D

Using the standard binomial distribution, with  $N = 12$ , we get

$$\begin{aligned} p &= \binom{12}{3} \left(\frac{1}{2}\right)^{12} + \binom{12}{2} \left(\frac{1}{2}\right)^{12} + \binom{12}{1} \left(\frac{1}{2}\right)^{12} + \binom{12}{0} \left(\frac{1}{2}\right)^{12} \\ &= 7.30\% \end{aligned}$$

**BUT...**

What if the experimenter had decided to stop when he had reached 3 D? Suddenly the sampling distribution is no longer the binomial distribution, but what is called the negative binomial distribution, and the values more “extreme” are different:

- 13 U + 3 D
- 14 U + 3 D
- 15 U + 3 D
- 16 U + 3 D
- ⋮

Numerically summing these terms (you can do it analytically, by looking at the first 11 terms in the negative binomial, and subtracting it from 1):

```
In [14]:p=0
In [15]:for N in range(12,100):
    .15.: p=p+nchoosek(N-1,3-1)*0.5**(N)
    .15.:
```

```
In [16]:p
Out[16]:0.03271484375
```

$$p = 3.27\%$$

In summary, if the experimenter **decided to flip 12 times**, then these results **would not reject** the null hypothesis of 50-50 chance at the 5% level, and this result or more extreme could “reasonably be expected to occur by chance if the pin was equally likely to fall in either position” (Lindley and Phillips, 1976).

If, however, the experimenter **decided to flip until there were 3 D**, then these results **would reject** the null hypothesis of 50-50 chance at the 5% level, and this result or more extreme could “not be reasonably be expected to occur by chance if the pin was equally likely to fall in either position” (Lindley and Phillips, 1976).

This isn’t the result of some small threshold difference, because the results are different by a factor of 2!

Now, what would happen if, as the Lindley states, that he was stopped when his wife finished making the coffee. What sampling distribution do you use then?

### 4.1.2 Bayesian Statistics

In the Bayesian approach, there is no stopping condition, and the mood of the experimenter plays no part in the analysis, because we are not comparing to data that wasn't measured, only the data we have. So we get, as in the proportion estimates above, that the median probability for a down tack is

```
In [17]:h=3
```

```
In [18]:N=12
```

```
In [19]:median=percentile_beta(h,N,0.5)
```

```
In [20]:median
```

```
Out[20]:array(0.27527583248615201)
```

(the mean is  $h/N = 0.25$ )

The confidence interval around the median is:

```
In [21]:c=0.95
```

```
In [22]:m1=percentile_beta(h,N,.5-c/2) # lower bound
```

```
In [23]:m2=percentile_beta(h,N,.5+c/2) # upper bound
```

```
In [24]:m1
```

```
Out[24]:array(0.090920394572096608)
```

```
In [25]:m2
```

```
Out[25]:array(0.53813153923404089)
```

The probability for the chance of D less than 50-50 is

```
In [26]:from scipy import stats
```

```
In [27]:stats.beta.cdf(.5,h+1,N-h+1)
```

```
Out[27]:array(0.953857421875)
```

which gives significance at the 5% level.

## 4.2 Type A Stars

The example given above in the  $\chi^2$  test has some peculiar properties. The example goes as follows (from (Loredo, 1990)). Suppose there is a prediction that the fraction of Type A stars is  $a = 0.1$  in a cluster. An astronomer measures 5 Type A stars out of 96, and is wondering whether the prediction is supported.

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(5 - 9.6)^2}{9.6} + \frac{(91 - 86.4)^2}{86.4} = 2.45\end{aligned}$$

The probability that  $\chi^2$  is greater than 2.45 is

```
In [31]:1-stats.chi2.cdf(2.45,1)
Out[31]:0.117524868097
```

So the prediction is acceptable (was not rejected) at the 95% confidence level.

If, however, the observer had decided to stop when he reached at 5 Type A stars, rather than 96 total, then the calculation would involve the negative binomial, and have gone like

```
In [32]:N=96; r=5; q=.9; p=.1; sigma=sqrt(r*q/p**2)
```

```
In [33]:sigma
Out[33]:21.2132034356
```

```
In [34]:mu=5/p
```

```
In [35]:mu
Out[35]:50.0
```

```
In [36]:chi2=(N-mu)**2/sigma**2
```

```
In [37]:chi2
Out[37]:4.70222222222
```

The probability that  $\chi^2$  is greater than 4.70 is

```
In [38]:1-stats.chi2.cdf(4.70,1)
Out[38]:0.0301626173985
```

So the hypothesis is rejected!

One Bayesian method using Jaynes'  $\psi$  measure gives

```
In [39]:n=array([5,91])
```

```
In [40]:p=array([.1,.9])
```

```
In [41]:psi=10*sum(n*log10(n/(96*p)))
```

```
In [42]:psi
Out[42]:6.33509992122
```

Another Bayesian method, like the coin flip methods above, give

```
In [43]:h=5; N=96
In [46]:1-stats.beta.cdf(.1,h+1,N-h+1)
Out[46]:0.0688994787386
```

Does not reject, but is not strong evidence for  $a = 0.1$ .

### 4.3 Cauchy Distribution

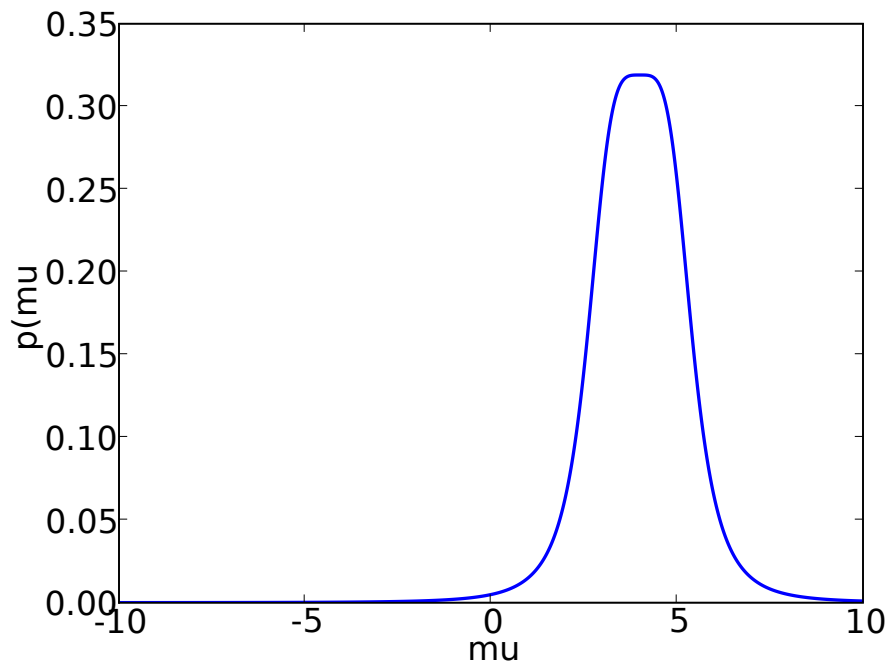
The Cauchy (sampling) distribution is

$$p(y|\mu) = \frac{1}{\pi} \left( \frac{1}{1 + (y - \mu)^2} \right)$$

The Bayesian posterior (with uniform prior) is

$$p(\mu|D, I) \propto \prod_{i=1}^n \frac{1}{1 + (\mu - y_i)^2}$$

For example, if  $N = 2$ ,  $y_1 = 3$ , and  $y_2 = 5$ , we have



Code:

```

from utils import *

data=[3,5] # data
n=len(data)
mu=r_[-10:10:.01]
dmu=mu[1]-mu[0]

p_mu=ones(mu.shape)

for y in data:
    p_mu=p_mu* (1/(1+(mu-y)**2))

p_mu=p_mu/sum(p_mu)/dmu

figure(1)
clf()
plot(mu,p_mu,linewidth=2)
ylabel('p(mu)')
xlabel('mu')
show()

savefig('fig080305_1.pdf')

```

### 4.3.1 Orthodox estimator?

One of the peculiar properties of this distribution is that the sampling distribution of the mean is

$$p(\bar{y}|\mu) \propto \frac{1}{1 + (\bar{y} - \mu)^2}$$

which is the same sampling distribution as a single data point! So, if the long-time sampling probabilities are the important ones to look at, then  $\bar{y}$ ,  $y_1$ , or any single  $y_i$ , should all be equivalent estimators. One should not prefer one over any other.

(Jaynes, 1976) goes through explicitly in the 2 data-point case.

# Chapter 5

## Misc

Here are some miscellaneous notes that I haven't had time to explore very well, and hope to get back to. Unfortunately, the summer is basically over, and I have to return my library books. :)

### 5.1 Max Entropy Derivations of Priors

These derivations are from (Sivia, 1996).

#### 5.1.1 Mean

Knowledge of mean

$$\langle x \rangle = \int xp(x|I)dx = \mu$$

Find maximum of

$$Q = - \sum_i p_i \log(p_i/m_i) + \lambda_0 \left(1 - \sum_i p_i\right) + \lambda_1 \left(\mu - \sum_i x_i p_i\right)$$

setting  $\partial Q/\partial p_j = 0$  we get

$$p_j = m_j e^{-(1+\lambda_0)} e^{-\lambda_1 x_j}$$

normalize from 0 to  $\infty$  we get

$$p(x|\mu) = \frac{1}{\mu} e^{-x/\mu}$$

#### 5.1.2 Mean and Second Moment

Knowledge of mean and second moment or, equivalently,  $\sigma^2$ .

$$\begin{aligned}\langle x \rangle &= \int xp(x|I)dx = \mu \\ \langle x^2 \rangle &= \int x^2p(x|I)dx \\ \langle (x - \mu)^2 \rangle &= \int (x - \mu)^2p(x|I)dx = \sigma^2\end{aligned}$$

Find maximum of

$$Q = - \sum_i p_i \log(p_i/m_i) + \lambda_o(1 - \sum_i p_i) + \lambda_1(\sigma^2 - \sum_i (x_i - \mu)^2 p_i)$$

leads to

$$p_j = m_j e^{-(1+\lambda_o)} e^{-\lambda_1(x_j - \mu)^2}$$

normalized gives

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

## 5.2 Derivation of Maximum Entropy

Derivation from (Jaynes, 1957).

Conditions for a measure of uncertainty of a distribution,  $H$ :

1.  $H$  is continuous function of  $p_i$
2. If all the  $p_i$  are equal, the quantity  $A(n) = H(1/n, 1/n, \dots, 1/n)$  is a monotonic increasing function of  $n$
3. Composition law. If instead of the  $p_i$  we group the first  $k$  as a single event with probability  $w_1 = p_1 + p_2 + \dots + p_k$  then the next  $m$  with  $w_2 = p_{k+1} + p_{k+2} + \dots + p_{k+m}$ , etc. Then we have uncertainty  $H(w_1, w_2, \dots, w_r)$ . If we are given the conditional probabilities  $(p_1/w_1, p_2/w_1, p_3/w_1, \dots, p_n/w_1)$  for the first event, and the same  $(p_1/w_2, p_2/w_2, p_3/w_2, \dots, p_n/w_2)$  for the second event, etc. Then we have arrived at the same state of knowledge as  $(p_1, p_2, p_3, \dots, p_n)$ . So we must get the same measure of uncertainty.

Thus we have

$$\begin{aligned}H(p_1, p_2, p_3, \dots, p_n) &= H(w_1, w_2, \dots, w_r) + \\ &w_1 H(p_1/w_1, p_2/w_1, p_3/w_1, \dots, p_n/w_1) + \\ &w_2 H(p_1/w_2, p_2/w_2, p_3/w_2, \dots, p_n/w_2) + \\ &\vdots \\ &w_r H(p_1/w_r, p_2/w_r, p_3/w_r, \dots, p_n/w_r)\end{aligned}$$

In words this is

*total uncertainty = grouped uncertainty + extra uncertainty due to event 1*, occurring with probability  $w_1$  + *uncertainty due to event 2*, occurring with probability  $w_2$  + etc. . .

For example,  $H(1/2, 1/3, 1/6) = H(1/2, 1/2) + \frac{1}{2}H(2/3, 1/3)$ .

From condition 1, it is enough to look at rational values

$$p_i = n_i / \sum_i n_i$$

where the  $n_i$  are integers.

From the composition law, we can look at them separately, or lumped altogether. For example, for  $\mathbf{n} = (3, 4, 2)$ :

$$A(9) = H\left(\frac{3}{9}, \frac{4}{9}, \frac{2}{9}\right) + \frac{3}{9}A(3) + \frac{4}{9}A(4) + \frac{2}{9}A(2)$$

In general, like

$$A\left(\sum_i n_i\right) = H(p_1, p_2, \dots, p_n) + \sum_i p_i A(n_i)$$

If all  $n_i = m$ , then this simplifies to

$$A(mn) = A(m) + A(n)$$

Which is solved by

$$A(n) = K \log n$$

Condition 2 requires  $K > 0$ .

Now we get

$$\begin{aligned} H(p_1, p_2, \dots, p_n) &= K \log \sum_i n_i - K \sum_i p_i \log n_i \\ &= -K \sum_i p_i \log p_i \end{aligned}$$

which is the form for entropy.

The continuous version is

$$H = - \int p(\theta) \log \left( \frac{p(\theta)}{m(\theta)} \right) d\theta$$

where  $m(\theta)$  is the least informative prior assignment for  $\theta$ .

## 5.3 Problem from Loredo

These sections are rough copies of sections from (Loredo, 1990).

### 5.3.1 Estimating the Amplitude of a Signal

Estimate the magnitude of a signal,  $\mu$ , for which there are  $N$  measurements  $x_i$  contaminated with noise with standard deviation  $\sigma$ .

#### 5.3.1.1 Frequentist Approach

In frequentist theory, the random variables are  $x_i$  not  $\mu$  (which is a constant parameter), each with a Gaussian distribution

$$p(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$

To estimate  $\mu$ , we choose a *statistic* – a function of the random variables – and calculate its distribution connecting it to  $\mu$ . One could choose, for example,  $x_3$  (the third sample),  $(x_1 + x_N)/2$  (the mean of the first and last samples), the median, or the sample mean,  $\bar{x} = \sum_i x_i/N$ . The criteria for choosing the “best” statistic are not unified. For example, one often wants an unbiased statistic (the long-run average equal to the real value), but *all* of the above are unbiased. Various criteria are used in this case, and the sample mean is chosen as the “best” estimate of  $\mu$ . How certain are we? We bring in the confidence region for  $\mu$ , from  $\bar{x}$ . The distribution of  $\bar{x}$  is calculated as

$$p(\bar{x}|\mu) = \left(\frac{N}{2\pi\sigma^2}\right)^{1/2} e^{-N(\bar{x}-\mu)^2/2\sigma^2}$$

Now, what confidence region do we take? The  $1\sigma$  region could be  $[-\infty, y]$ ,  $[-\sigma/\sqrt{N}, +\sigma/\sqrt{N}]$ , or  $[-y, \infty]$ . One usually chooses (what basis?) the shortest confidence interval, or the symmetric one (which is the same in this case).

#### 5.3.1.2 Bayesian Approach

This analysis is the same as in Section 2.1, with a proper prior on  $\mu$ .

$$p(\mu|\sigma, I) = p(\mu|I) = \begin{cases} \frac{1}{\mu_{\max}-\mu_{\min}} & \mu_{\min} \leq \mu \leq \mu_{\max} \\ 0 & \text{otherwise} \end{cases}$$

The likelihood is identical to that done before

$$p(\mathbf{x}|\mu, \sigma, I) = \prod_{k=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_k-\mu)^2/2\sigma^2}$$

We then get the posterior as

$$p(\mu|\mathbf{x}, \sigma, I) = \left[ \frac{1}{2} \operatorname{erf} \left( \frac{\bar{x} - \mu_{\max}}{\sqrt{2N}\sigma_2} \right) - \frac{1}{2} \operatorname{erf} \left( \frac{\bar{x} - \mu_{\min}}{\sqrt{2N}\sigma_2} \right) \right] \left( \frac{N}{2\pi\sigma^2} \right)^{1/2} e^{-N(\bar{x}-\mu)^2/2\sigma^2}$$

In the limit that  $\mu_{\max}$  and  $\mu_{\min}$  are far apart, we get the result from Section 2.1. As before, we can estimate mean values, median values, highest probability intervals, etc.

### 5.3.1.3 Comparison

To further quote (Loredo, 1990):

To a Bayesian,  $\bar{x}$  is the most plausible value of  $\mu$  given the one set of data at hand, and there is a plausibility of 0.68 that  $\mu$  is in the range  $\bar{x} \pm \sigma/\sqrt{N}$ . In contrast, the frequentist interpretation of the result is a statement about the long term performance of adopting the procedure of estimating  $\mu$  with  $\bar{x}$  and stating that the true value of  $\mu$  is in the interval  $\bar{x} \pm \sigma/\sqrt{N}$ . Specifically, if one adopts this procedure, the average of the estimates of  $\mu$  after many observations will converge to the true value of  $\mu$ , and the statement about the interval containing  $\mu$  will be true 68% of the time.

The frequentist procedure has to appeal to *ad hoc* criteria, like unbiasedness and shortest confidence intervals, which are not universally correct. The Bayesian procedure gives a unique solution to well-posed problems, and is **guaranteed to be the best one possible given the information, given our criteria of rationality and consistency.**

## 5.3.2 Measuring a Weak Counting Signal

We want to look at the measurement of a signal in the presence of a background rate. The usual approach is to estimate the background rate,  $\hat{b}$ , and its standard deviation,  $\sigma_b$ , by observing an empty part of the sky, and an estimate of the signal plus the background rate,  $\hat{r}$ , and its standard deviation,  $\sigma_r$ , by observing the region where the signal is expected. The signal rate is then estimated by  $\hat{s} = \hat{r} - \hat{b}$ , with variance  $\sigma_s^2 = \sigma_r^2 + \sigma_b^2$ . This is correct, for Gaussian distributions, but can fail when we have other distributions. It can lead to negative estimates of the signal strength, or the confidence regions can include negative values. This is particularly a problem with low  $N$ , as in high-energy astronomy. Three frequentist alternatives have been proposed for gamma-ray astronomers (Hearn 1969; O'Mongain 1973; Cherry et. al. 1980), none of which takes into account the uncertainty in the background rate. The problems are due to the fact that there is a nuisance parameter (the background rate) and there is prior information that neither the signal nor the background can be negative.

The solution is as follows. The background rate,  $b$ , is measured with  $n_b$  events in time  $T$  from an “empty” part of the sky. From the Poisson distribution we get the likelihood

$$p(n_b|bI_b) = \frac{(bT)_b^n e^{-bT}}{n_b!}$$

The least informative prior, from Jaynes' invariance arguments, is the Jeffreys prior:

$$p(b|I_b) = \frac{1}{b}$$

with prior probabilities for  $b < 0$  set to 0. The marginal likelihood is

$$\begin{aligned} p(n_b|I_b) &= \int db p(n_b|bI_b)p(b|I_b) \\ &= \frac{T_b^n}{n_b!} \int_0^\infty b^{n_b-1} e^{-bT} db \\ &= \frac{1}{n_b} \end{aligned}$$

The posterior density for  $b$  is

$$p(b|n_bI_b) = \frac{T(bT)^{n_b-1} e^{-bT}}{n_b - 1}$$

The average background rate is calculated  $\langle b \rangle = n_b/T$  with standard deviation  $n_b^{(1/2)}/T$ . If we count  $n$  events in time  $t$  from the part of the sky with the suspected signal, we get

$$\begin{aligned} p(n|sbI) &= \frac{t^n (s+b)^n e^{-(s+b)t}}{n!} \\ p(s|bI) &= \frac{1}{s+b} \end{aligned}$$

again, with the prior truncated at zero.

$$\begin{aligned} p(sb|nI) &= p(s|bI)p(b|I) \frac{p(n|sbI)}{p(n|I)} \\ &\propto (s+b)^{n-1} b^{n_b-1} e^{-st} e^{-b(t+T)} \end{aligned}$$

Marginalizing with respect to  $b$

$$\begin{aligned} p(s|nI) &= \int p(sb|nI) db \\ &= \sum_{i=1}^n C_i \frac{t(st)^{i-1} e^{-st}}{(i-1)!} \end{aligned}$$

with

$$C_i \equiv \frac{(1 + \frac{T}{t})^i \frac{(n+n_b-i-1)!}{(n-i)!}}{\sum_j (1 + \frac{T}{t})^j \frac{(n+n_b-j-1)!}{(n-j)!}}$$

Although a bit ugly, it is a straightforward calculation.

## 5.4 Anova and T-distributions

This info is coming from (Lee, 1989), pages 185-191. I don't understand his notation much of the time, and he uses variance as opposed to standard deviation as the variable of choice for

distribution width, so watch out! He also is loose about not putting the background information conditional symbol. I hope to clear this up in my head later.

In comparing more than one sample, we have a vector of unknown parameters,  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_I)$ , with  $N = \sum K_i$  independent observations of each

$$x_{ik} \sim N(\theta_i, \phi) \quad (i = 1, 2, \dots, I; k = 1, 2, \dots, K_i)$$

with a common variance,  $\phi$ .

$$p(\boldsymbol{\theta}, \phi) \propto 1/\phi \text{ (prior)}$$

The likelihood is

$$(2\pi\phi)^{-N/2} \exp(-\frac{1}{2}S/\phi)$$

where

$$S = \sum_i \sum_k (x_{ik} - \theta_i)^2$$

so the posterior is

$$p(\boldsymbol{\theta}, \phi | \mathbf{x}) \propto \phi^{-N/2-1} \exp(-\frac{1}{2}S/\phi)$$

Define

$$\begin{aligned} x_{i.} &\equiv \sum_k x_{ik}/K_i \\ x_{..} &\equiv \sum_i \sum_k x_{ik}/N \\ \lambda &\equiv \sum_i \theta_i/I \\ \mu_i &\equiv \theta_i - \lambda \\ \hat{\lambda} &\equiv x_{..} \\ \hat{\mu}_i &\equiv x_{i.} - x_{..} \end{aligned}$$

We will often want to test whether then  $\theta_i$  are all 0, or equivalently, whether  $\mu_1 = \mu_2 = \dots = \mu_I = 0$ .

The Jacobian from  $\boldsymbol{\theta}, \phi$  to  $\lambda, \mu_1, \mu_2, \dots, \mu_{I-1}, \phi$  is constant. (The  $\mu_I$  is constrained to satisfy  $\sum_i K_i \mu_i = 0$ ). So

$$p(\lambda, \boldsymbol{\mu}, \phi | \mathbf{x}) = p(\boldsymbol{\theta}, \phi | \mathbf{x})$$

Reexpressing  $S$  in terms of  $\lambda, \boldsymbol{\mu}, \phi$ . Since  $x_{i.} = \hat{\lambda} + \hat{\mu}_i$  and  $\theta_i = \lambda + \mu_i$  it follows that

$$-(x_{ik} - \theta_i) = (\lambda - \hat{\lambda}) + (\mu_i - \hat{\mu}_i) + (x_{i.} - x_{ik})$$

it then follows

$$\begin{aligned}
 S &= \sum \sum (x_{ik} - \theta_i)^2 \\
 &= N(\lambda - \hat{\lambda})^2 + S_t(\boldsymbol{\mu}) + S_e \\
 S_t(\boldsymbol{\mu}) &= \sum K_i(\mu_i - \hat{\mu})^2 \\
 S_e &= \sum \sum (x_{i.} - x_{ik})^2
 \end{aligned}$$

Define  $\nu = N - I$  and  $s^2 = S_e/\nu$ . The posterior is now

$$p(\lambda, \boldsymbol{\mu}, \phi | \mathbf{x}) \propto \phi^{-N/2-1} \exp \left[ -(N(\lambda - \hat{\lambda})^2 + S_t(\boldsymbol{\mu}) + S_e)/2\phi \right]$$

Integrate out  $\lambda$ .

$$p(\boldsymbol{\mu}, \phi | \mathbf{x}) \propto \phi^{-N/2-1} \exp \left[ -(S_t(\boldsymbol{\mu}) + S_e)/2\phi \right]$$

Integrate out  $\phi$ .

$$\begin{aligned}
 p(\boldsymbol{\mu} | \mathbf{x}) &\propto (S_t(\boldsymbol{\mu}) + S_e)^{-(N-1)/2} \\
 &\propto (1 + (I-1)F(\boldsymbol{\mu})/\nu) \phi^{-(N-1)/2}
 \end{aligned}$$

where

$$F(\boldsymbol{\mu}) = \frac{S_t(\boldsymbol{\mu})/(I-1)}{S_e/\nu} = \frac{\sum K_i(\mu_i - \hat{\mu})^2/(I-1)}{s^2}$$

The posterior for  $\boldsymbol{\mu}$  is the *multivariate t-distribution*. Has a maximum at  $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$ . HDR region is of the form

$$E(F) = \{\boldsymbol{\mu}; F(\boldsymbol{\mu}) \leq F\}$$

and

$$F(\boldsymbol{\mu}) \sim F_{I-1, \nu}$$

For one-way anova, this is the same as the significance test for  $\boldsymbol{\mu} = \mathbf{0}$ .

Anova table constructed as follows

$$\begin{aligned}
 S_t(\mathbf{0}) &= \sum K_i(x_{i.} - x_{..})^2 \equiv S_t \\
 S_T &= \sum \sum (x_{ik} - x_{..})^2 \\
 &= S_t + S_e
 \end{aligned}$$

Subtract a useless constant

$$\begin{aligned}
 S_T &= \sum \sum x_{ik}^2 - Nx_{..}^2 = \sum x_{ik}^2 - C \\
 S_t &= \sum K_i x_{i.}^2 - Nx_{i.}^2 = \sum T_i^2 / K_i - C \\
 T &\equiv \sum x_{ik} = K_i x_{i.} \text{ (total for treatment } i) \\
 G &\equiv \sum \sum x_{ik} = Nx_{..} \text{ (grand total)} \\
 C &= G^2 / N \text{ ("correction for error")}
 \end{aligned}$$

Then we find  $F(\mathbf{0})$  by making the table

Source	Sum of Squares	Deg of Freedom	Mean Square	Ratio
Treatments	$S_t$	$I - 1$	$S_t / (I - 1)$	$F(\mathbf{0})$
Error	$S_e$	$\nu = N - 1$	$s^2 = S_e / \nu$	
Total	$S_T$	$N - 1$		

Example from an experiment on the effect of sulphur in reducing scab disease in potatoes. In addition to the untreated plots which serve as a control, three amounts of dressing were compared—300, 600, and 1200 pounds per acre. Both an autumn and a spring application of each treatment were tried, do that in all there were seven distinct treatments. The effectiveness of the treatments was measured by the “scab index”, which is (roughly) the average percentage of the area of 100 potatoes taken at random from each plot that is affected with scab. The data are as follows

$i$	Treatment	$K_i$	Scab Indices $x_{ik}$								$T_i$	$x_{i.}$	$\hat{\mu}_i$
1	0	8	12	30	10	18	24	32	29	26	181	22.6	7.0
2	A3	4	9	9	16	4					38	9.5	-6.2
3	S3	4	30	7	21	9					67	16.8	1.1
4	A6	4	16	10	18	18					62	15.5	-0.2
5	S6	4	18	24	12	19					73	18.2	2.6
6	A12	4	10	4	4	5					23	5.8	-9.9
7	S12	4	17	7	16	17					57	14.2	-1.2

There are  $I = 7$  treatments and  $N = \sum K_i = 32$  observations, the grand total being  $G = 501$  (and the grand average  $x_{..} = 15.66$ ), the crude sum of squares  $\sum \sum x_{ik}^2 = 9939$  and the correction for error  $C = G^2 / N = 7844$ . Also,  $S_T = 9939 - 7844 = 2095$  and  $S_t = 181^2 / 8 + (38^2 + 67^2 + 62^2 + 73^2 + 23^2 + 57^2) / 4 - 7844 = 972$ .

The table is now

Source	Sum of Squares	Deg of Freedom	Mean Square	Ratio
Treatments	972	6	162	3.60
Error	1123	25	45	
Total	2095	31		

From the tables of the F distribution, an  $F_{6,25}$  variable exceeds 3.63 with probability 0.01.

# Appendix A

## Supplementary Code

### A.1 `utils.py`

```
from pylab import *
from numpy import *
import bigfonts
bigfonts.bigfonts()

from scipy.stats import beta

from matplotlib.mlab import hist

def get_data(dataset):

    fname='../Business Statistics in Practice 3rd Edition/Bowerman_Stuff/Datasets - Text
    fname=fname+dataset+'.txt'

    if dataset=='WaitTime' or dataset=='CustSat':

        data=[]

        for line in open(fname).readlines():
            parts=line.split()
            try:
                data.append(float(parts[0]))
            except ValueError: # first line
                pass

        return array(data)
```

```
    else:
        raise ValueError, "Unimplemented dataset %s" % dataset

def percentile_beta(h,N,p):

    a=h+1
    b=(N-h)+1

    y=beta.ppf(p,a,b)

    return y

def factorial(N):

    y=1
    for i in range(1,N+1):
        y=y*i

    return y

def nchoosek(N,k):

    y=factorial(N)/factorial(k)/factorial(N-k)

    return y
```

# Appendix B

## Derivations for Single Samples

### B.1 Unknown $\mu$ , Known $\sigma$

This derivation is taken directly from (Sivia, 1996), with some of the steps filled in and elaborated by me.

#### B.1.1 Introducing the Distributions

We want to estimate the mean of data, given the standard deviation. Here we want

$$p(\mu|\mathbf{x}, \sigma, I)$$

where  $\mu$  is the parameter we want to estimate,  $x_k$  are the data (the full vector abbreviated as  $\mathbf{x}$ ),  $\sigma$  is the (known) standard deviation of the distribution, and  $I$  is any other background information. In other words, we would like the probability of the parameter ( $\mu$ ) given the data ( $\mathbf{x}$ ). From Bayes' theorem we get the *posterior* probability distribution for  $\mu$ :

$$p(\mu|\mathbf{x}, \sigma, I) \propto p(\mathbf{x}|\mu, \sigma, I) \cdot p(\mu|\sigma, I)$$

The *prior* information,  $p(\mu|\sigma, I)$ , about the value we want to estimate,  $\mu$ , will be as non-informative as possible (we are really ignorant of its value). Thus, we use a uniform distribution:

$$p(\mu|\sigma, I) = p(\mu|I) = \begin{cases} A & \mu_{\min} \leq \mu \leq \mu_{\max} \\ 0 & \text{otherwise} \end{cases}$$

The *likelihood*, also known as the generative probability because it is the probability that the data would be generated from the model, is

$$p(\mathbf{x}|\mu, \sigma, I)$$

If we assume *independent, Gaussian* distributions<sup>1</sup> for each data point, with a *known*  $\sigma$ , then we have

$$p(x_k|\mu, \sigma, I) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_k-\mu)^2/2\sigma^2}$$

---

<sup>1</sup>For a full justification of Gaussian assumptions, please read Jaynes Chapter ??.

$$p(\mathbf{x}|\mu, \sigma, I) = \prod_{k=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_k - \mu)^2/2\sigma^2}$$

### B.1.2 An Aside on Log Posterior

It is often convenient to use the log of the posterior distribution, rather than the posterior distribution itself. It has the same maxima, but is easier to analyze in many cases, especially with Gaussian assumptions.

If, for example, the posterior has a Gaussian form:

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-y^2/2\sigma^2}$$

then the log posterior is

$$L(y) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{y^2}{2\sigma^2}$$

The maximum is found by setting the first derivative to zero, getting the obvious result.

$$\begin{aligned} \frac{dL(y)}{dy} &= -y/\sigma^2 = 0 \\ \Rightarrow y &= 0 \end{aligned}$$

If we look at the second derivative, we get

$$\begin{aligned} \frac{d^2L(y)}{dy^2} &= -\frac{1}{\sigma^2} \\ \Rightarrow \sigma &= \left(-\frac{d^2L(y)}{dy^2}\right)^{-1/2} \end{aligned}$$

So, under the Gaussian posterior assumption, then the width of the posterior distribution is related to the second derivative of the log posterior.

### B.1.3 Continuing

Now we are in a position to calculate the probability distribution for our parameter,  $\mu$ , obtain the best estimate and the confidence intervals.

$$\begin{aligned} p(\mu|\mathbf{x}, \sigma, I) &\propto p(\{x_k\}|\mu, \sigma, I) \cdot p(\mu|\sigma, I) \\ &\propto \prod_{k=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_k - \mu)^2/2\sigma^2} \\ L(\mu|x_1, x_2, \dots, x_n, \sigma, I) &= \text{constant} - \sum_{k=1}^N (x_k - \mu)^2/2\sigma^2 \end{aligned}$$

The best estimate,  $\hat{\mu}$ , is that which maximizes  $L$

$$\begin{aligned}\frac{dL}{d\mu}\Big|_{\hat{\mu}} &= +\frac{1}{\sigma^2} \sum_{k=1}^N (x_k - \mu)\Big|_{\hat{\mu}} = 0 \\ \sum_{k=1}^N x_k - \sum_{k=1}^N \hat{\mu} &= 0 \\ \hat{\mu} &= \frac{\sum_{k=1}^N x_k}{N}\end{aligned}$$

which is just the sample mean. If we denote 1 standard deviation confidence intervals in our estimate of  $\mu$ , then our best estimate is  $\mu = \hat{\mu} \pm \sigma_{\mu}$ . The width of the distribution, which gives us the confidence interval, is related to the second derivative of  $L$ .

$$\begin{aligned}\sigma_{\mu} &= \left( -\frac{d^2L(y)}{d\mu^2}\Big|_{\hat{\mu}} \right)^{-1/2} \\ \frac{d^2L(y)}{d\mu^2}\Big|_{\hat{\mu}} &= -\frac{N}{\sigma^2} \\ \sigma_{\mu} &= \frac{\sigma}{\sqrt{N}}\end{aligned}$$

In summary, our best estimate of the value of a parameter,  $\mu$ , given the data  $\{x_k\}$  and the standard deviation of the likelihood,  $\sigma$ , is

$$\begin{aligned}\mu &= \frac{\sum_{k=1}^N x_k}{N} \pm \frac{\sigma}{\sqrt{N}} \\ &\equiv \bar{x} \pm \frac{\sigma}{\sqrt{N}}\end{aligned}\tag{B.1.1}$$

Our posterior probability distribution for  $\mu$  is

$$p(\mu|\mathbf{x}, \sigma, I) = \sqrt{\frac{N}{2\pi\sigma^2}} e^{-N(\bar{x}-\mu)^2/2\sigma^2}\tag{B.1.2}$$

## B.2 Unknown $\mu$ , Unknown $\sigma$

This derivation is taken directly from (Sivia, 1996), with some of the steps filled in and elaborated by me.

### B.2.1 Setting up the Problem

If both  $\mu$  and  $\sigma$  are unknown, then the procedure is as follows. We will need the joint distribution of both variables

$$p(\mu, \sigma | \mathbf{x}, I)$$

and then integrate the “nuisance” parameter, to get the posterior for the parameter we are interested in. If we are estimating the mean, we need to integrate out the standard deviation, and vice versa:

$$\begin{aligned} p(\mu | \mathbf{x}, I) &= \int_0^\infty p(\mu, \sigma | \mathbf{x}, I) d\sigma \\ p(\sigma | \mathbf{x}, I) &= \int_{-\infty}^\infty p(\mu, \sigma | \mathbf{x}, I) d\mu \end{aligned}$$

The joint distribution can be found using Bayes rule, Gaussian likelihoods, and simple flat priors. We will see how using the correct (Jeffrey’s) prior for  $\sigma$  will slightly modify our results.

$$\begin{aligned} p(\mu, \sigma | \mathbf{x}, I) &\propto p(\mathbf{x} | \mu, \sigma, I) p(\mu, \sigma | I) \\ p(\mathbf{x} | \mu, \sigma, I) &= \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^N e^{-\frac{1}{2\sigma^2} \sum_{k=1}^N (x_k - \mu)^2} \\ p(\mu, \sigma | I) &= \begin{cases} \text{constant} & \sigma > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Plugging into the posterior for the mean we get

$$\begin{aligned} p(\mu | \mathbf{x}, I) &= \int_0^\infty p(\mu, \sigma | \mathbf{x}, I) d\sigma \\ &\propto \int_0^\infty d\sigma \frac{1}{\sigma^N} e^{-\frac{1}{2\sigma^2} \sum_{k=1}^N (x_k - \mu)^2} \end{aligned}$$

For convenience, make the substitution  $\sigma = 1/t$ , which implies  $d\sigma = -dt/t^2$ , and we have

$$\begin{aligned} p(\mu | \mathbf{x}, I) &\propto \int_0^\infty t^{N-2} e^{\frac{t^2}{2} \sum_{k=1}^N (x_k - \mu)^2} dt \\ &\propto \left[ \sum_{k=1}^N (x_k - \mu)^2 \right]^{(N-1)/2} \end{aligned}$$

The last step of the analysis was obtained by noticing that

$$\int_0^\infty t^{N-2} e^{\frac{t^2}{2} \sum_{k=1}^N (x_k - \mu)^2} dt$$

is of the form of a simple power ( $t^{N-2}$ ) multiplied by a Gaussian:

$$\int_0^\infty t^n e^{-at^2} dt \equiv I_n$$

and the Gaussian integral tricks listed later (Appendix E.4) are then used.

## B.2.2 Estimating the Mean

We finished the previous section with the tools to determine the posterior for the mean:

$$\begin{aligned} p(\mu|\mathbf{x}, I) &\propto \int_0^\infty t^{N-2} e^{\frac{t^2}{2} \sum_{k=1}^N (x_k - \mu)^2} dt \\ &\propto \left[ \sum_{k=1}^N (x_k - \mu)^2 \right]^{(N-1)/2} \end{aligned}$$

Now, with the standard trick used before (assuming a Gaussian posterior distribution), we look at the derivatives of the log-posterior. Setting the first derivative to zero gives us our estimate,

$$\begin{aligned} L(\mu) &= \text{constant} - \frac{(N-1)}{2} \log \sum_{k=1}^N (x_k - \mu)^2 \\ \frac{dL}{d\mu} &= \frac{(N-1) \sum_{k=1}^N (x_k - \mu)}{\sum_{k=1}^N (x_k - \mu)^2} = 0 \\ &= \frac{(N-1) \left( \sum_{k=1}^N x_k - N\mu \right)}{\sum_{k=1}^N (x_k - \mu)^2} = 0 \\ \Rightarrow \hat{\mu} &= \frac{1}{N} \sum_{k=1}^N x_k \end{aligned}$$

Calculating the second derivative allows us to determine confidence intervals.

$$\begin{aligned} \left. \frac{d^2 L}{d\mu^2} \right|_{\hat{\mu}} &= - \left. \frac{N(N-1)}{\sum_{k=1}^N (x_k - \mu)^2} \right|_{\hat{\mu}} + \underbrace{\left. \frac{(N-1) \sum_{k=1}^N (x_k - \mu) \cdot 2 \sum_{k=1}^N (x_k - \mu)}{\sum_{k=1}^N (x_k - \mu)^3} \right|_{\hat{\mu}}}_{=0} \\ \text{width} &= - \left[ \left. \frac{d^2 L}{d\mu^2} \right|_{\hat{\mu}} \right]^{-1/2} \\ &= \frac{1}{\sqrt{N(N-1)}} \sum_{k=1}^N (x_k - \mu) \end{aligned}$$

So our final estimate is

$$\begin{aligned} \mu &= \bar{x} \pm \frac{S}{\sqrt{N}} \\ \bar{x} &= \frac{1}{N} \sum_{k=1}^N x_k \\ S^2 &= \frac{1}{(N-1)} \sum_{k=1}^N (x_k - \mu)^2 \end{aligned}$$

### B.2.3 A More Convenient Form

If we rewrite the sum  $\sum_{k=1}^N (x_k - \mu)^2$  as

$$\begin{aligned}\sum_{k=1}^N (x_k - \mu)^2 &\equiv N(\bar{x} - \mu)^2 + V \\ V &\equiv \sum_{k=1}^N (x_k - \bar{x})^2\end{aligned}$$

Then the posterior is

$$p(\mu|\mathbf{x}, I) \propto [N(\bar{x} - \mu)^2 + V]^{(N-1)/2}$$

which is the Student-t distribution with  $N-2$  degrees of freedom. Had we used the suggested (Jeffreys) prior for the scale parameter,  $\sigma$ ,

$$p(\mu, \sigma|I) = \begin{cases} \frac{1}{\sigma} & \sigma > 0 \\ 0 & \text{otherwise} \end{cases}$$

which simply adds a factor  $t$  to the posterior integral, and the result is

$$p(\mu|\mathbf{x}, I) \propto [N(\bar{x} - \mu)^2 + V]^{N/2}$$

again, the Student-t distribution but with  $N-1$  degrees of freedom.

We now define

$$t \equiv \frac{\mu - \bar{x}}{S/\sqrt{N}}$$

This variable  $t$  has exactly the standard t-distribution with  $N-1$  degrees of freedom, in a form which can be readily looked up.

### B.2.4 Estimating $\sigma$

With the Jeffrey's prior,

$$\begin{aligned}p(\sigma|\mathbf{x}, I) &= \int_{-\infty}^{\infty} p(\mu, \sigma|\mathbf{x}, I) d\mu \\ &\propto \frac{1}{\sigma^{N+1}} e^{-V/2\sigma^2} \underbrace{\int_{-\infty}^{\infty} e^{-\frac{N(\bar{x}-\mu)^2}{2\sigma^2}} d\mu}_{\propto \sigma} \\ &\propto \frac{1}{\sigma^N} e^{-V/2\sigma^2}\end{aligned}$$

If we change variables to  $\xi \equiv V/\sigma^2$ , then this is the  $\chi^2$  distribution with  $f \equiv N - 1$  degrees of freedom:

$$\begin{aligned} d\sigma &\propto \xi^{-3/2} \\ p(\xi|\mathbf{x}, I) &= p(\sigma|\mathbf{x}, I) \times \left| \frac{d\sigma}{d\xi} \right| \\ &\propto \xi^{N/2} e^{-\xi/2} \times \xi^{-3/2} \\ &\propto \xi^{N/2-1} e^{-\xi/2} \\ &\propto \xi^{f/2-1} e^{-\xi/2} \end{aligned}$$

If we used the uniform prior, then this would be  $f = N - 2$  degrees of freedom. With the uniform prior,

$$\begin{aligned} p(\sigma|\mathbf{x}, I) &= \int_{-\infty}^{\infty} p(\mu, \sigma|\mathbf{x}, I) d\mu \\ &\propto \frac{1}{\sigma^N} e^{-V/2\sigma^2} \underbrace{\int_{-\infty}^{\infty} e^{-\frac{N(\bar{x}-\mu)^2}{2\sigma^2}} d\mu}_{\propto \sigma} \\ &\propto \frac{1}{\sigma^{N-1}} e^{-V/2\sigma^2} \end{aligned}$$

$$\begin{aligned} L(\sigma) &= \text{constant} - (N-1) \log \sigma - \frac{V}{2\sigma^2} \\ \frac{dL}{d\sigma} &= -\frac{N-1}{\sigma} + \frac{V}{\sigma^3} = 0 \\ \Rightarrow \hat{\sigma}^2 &= \frac{V}{N-1} \\ \left. \frac{d^2L}{d\sigma^2} \right|_{\hat{\sigma}} &= \frac{N-1}{\hat{\sigma}^2} - \frac{3V}{\hat{\sigma}^4} \\ &= \frac{N-1}{\hat{\sigma}^2} - \frac{3(N-1)}{\hat{\sigma}^2} \\ &= -\frac{2(N-1)}{\hat{\sigma}^2} \\ \text{width} &= -\left[ \left. \frac{d^2L}{d\sigma^2} \right|_{\hat{\sigma}} \right]^{-1/2} \\ &= \frac{\hat{\sigma}}{\sqrt{2(N-1)}} \end{aligned}$$

So our final estimate for the mean and standard deviation is

$$\mu = \bar{x} \pm \frac{S}{\sqrt{N}}$$

$$\begin{aligned}\bar{x} &= \frac{1}{N} \sum_{k=1}^N x_k \\ S^2 &= \frac{1}{(N-1)} \sum_{k=1}^N (x_k - \bar{x})^2 \\ \sigma &= S^2 \pm \frac{S^2}{\sqrt{2(N-1)}}\end{aligned}$$

The possible negative values are an indication that the Gaussian approximation breaks down, and the actual pdf should be used.

### B.3 Unknown proportion

If  $\theta$  is the model representing the probability,  $\theta$ , of the coin landing on heads (and  $1 - \theta$  is the probability of landing on tails), we need to make an estimate of probability of model  $\theta$  being true given the data, which will consist of  $N$  flips of which  $h$  are heads.

Bayes rule is:

$$p(\theta|D, I) = \frac{p(D|\theta, I)p(\theta|I)}{p(D|I)} = \frac{p(D|\theta, I)p(\theta, I)}{\sum_{\theta} p(D|\theta, I)p(\theta|I)}$$

Thus, the probability of a particular model  $\theta$  being true is the product of the probability of the observed data ( $h$  heads in  $N$  flips) given the model  $\theta$  and the prior probability of the model  $\theta$  being true before we even look at the data, divided by the probability of the data itself over all models.

The prior probability of model  $\theta$  will be assumed to be uniform (from maximum entropy considerations). The probability,  $\theta$ , ranges from 0 to 1, to the prior is

$$p(\theta|I) = 1$$

The probability of the data given the random model, is just the binomial distribution:

$$p(D|\theta) = \binom{N}{h} \theta^h (1 - \theta)^{N-h}$$

The probability of the data,  $p(D|I)$ , is found by summing (or in this case integrating)  $p(D|\theta, I)p(\theta|I)$  for all  $\theta$ :

$$\begin{aligned}p(D|I) &= \int_0^1 \binom{N}{h} \theta^h (1 - \theta)^{N-h} \cdot 1 d\theta \\ &= \frac{N!}{h!(N-h)!} \frac{h!(N-h)!}{(N+1)!} = \frac{1}{N+1}\end{aligned}$$

Now the probability of model  $\theta$  being true, given the data, is just

$$\begin{aligned}
p(\theta|D, I) &= (N+1) \cdot \binom{N}{h} \theta^h (1-\theta)^{N-h} \\
&= \frac{(N+1)!}{h!(N-h)!} \theta^h (1-\theta)^{N-h}
\end{aligned}$$

### B.3.1 Max, Mean, Variance

The model with the maximum probability is found by maximizing  $p(\theta|D, I)$  w.r.t.  $\theta$ :

$$\begin{aligned}
\frac{dP(\theta|D, I)}{d\theta} &= 0 = \frac{(N+1)!}{h!(N-h)!} \left( -(N-h)\theta^h(1-\theta)^{N-h-1} + h\theta^{h-1}(1-\theta)^{N-h} \right) \\
(N-h)\theta^h(1-\theta)^{N-h-1} &= h\theta^{h-1}(1-\theta)^{N-h} \\
\theta(N-h) &= (1-\theta)h = h - \theta h = N\theta - \theta h \\
\theta &= \frac{h}{N} \quad \checkmark
\end{aligned}$$

The average and the standard deviation is also straightforward.

$$\begin{aligned}
\bar{\theta} &= \int_0^1 \theta \cdot \frac{(N+1)!}{h!(N-h)!} \theta^h (1-\theta)^{N-h} \\
&= \frac{(N+1)!}{h!(N-h)!} \int_0^1 \theta^{h+1} (1-\theta)^{N-h} \\
&= \frac{(N+1)!}{h!(N-h)!} \frac{(h+1)!(N-h)!}{(N+2)!} \\
&= \frac{h+1}{N+2} \\
\bar{\theta}^2 &= \int_0^1 \theta^2 \cdot \frac{(N+1)!}{h!(N-h)!} \theta^h (1-\theta)^{N-h} \\
&= \frac{(N+1)!}{h!(N-h)!} \frac{(h+2)!(N-h)!}{(N+3)!} \\
&= \frac{(h+1)(h+2)}{(N+2)(N+3)} \\
\sigma^2 &= \bar{\theta}^2 - \bar{\theta}^2 = \frac{(h+1)(h+2)}{(N+2)(N+3)} - \frac{(h+1)(h+1)}{(N+2)(N+2)} \\
&= \frac{(h+1)(N-h+1)}{(N+2)^2(N+3)} \\
&= \frac{(h+1)}{(N+2)} \left( \frac{N+2}{N+2} - \frac{h+1}{N+2} \right) \frac{1}{N+3} \\
&= \bar{\theta}(1-\bar{\theta}) \frac{1}{N+3}
\end{aligned}$$

**B.3.1.1 An Approximation for the Variance**

If  $f = h/N$  is the actual fraction of heads observed, then the variance above can be written as

$$\begin{aligned}\sigma^2 &= \frac{(fN + 1)(N - fN + 1)}{(N + 2)^2(N + 3)} \\ (\text{for large } N) &\approx \frac{(fN + 1)(N - fN)}{N^3} = \frac{(fN + 1)(1 - f)}{N^2} \\ (\text{for large } fN) &\approx \frac{(fN)(N - fN)}{N^2} = \frac{f(1 - f)}{N} \\ \sigma^2 &\approx \frac{f(1 - f)}{N}\end{aligned}$$

In this limit, the distribution (beta distribution) can be approximated with a Gaussian.

# Appendix C

## Derivations for Two Samples

### C.1 Paired Data Difference of Means, $\delta_k \equiv x_k - y_k$

We want

$$p(\mu_\delta | \mathbf{x}, \mathbf{y}, \sigma_x, \sigma_y, I)$$

where  $\delta_k \equiv x_k - y_k$ .

We have from Section B.1 the following likelihoods for  $x_k$  and  $y_k$ :

$$\begin{aligned} p(x_k | \mu, \sigma_x, I) &= \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-(x_k - \mu_x)^2 / 2\sigma_x^2} \\ p(y_k | \mu, \sigma_y, I) &= \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-(y_k - \mu_y)^2 / 2\sigma_y^2} \end{aligned}$$

Now we need to find the likelihood function for  $\delta_k \equiv x_k - y_k$ .

#### C.1.1 Changing Variables

If we have  $Z = f(X, Y)$ , and we know about  $X$  and  $Y$ , we can learn about  $Z$ .

$$\begin{aligned} p(Z|I) &= \int \int p(Z|X, Y, I) \times p(X, Y|I) dX dY \\ &= \int \int \delta(Z - f(X, Y)) \times p(X, Y|I) dX dY \end{aligned}$$

Say,  $Z = X - Y$ , and  $X$  and  $Y$  are independent, then  $p(X, Y|I) = p(X|I)p(Y|I)$  and we have

$$\begin{aligned} p(Z|I) &= \int dX p(X, I) \int dY p(Y|I) \delta(Z - X + Y) \\ &= \int dX p(X, I) p(Y = X - Z|I) \end{aligned}$$

Further, if the probabilities are Gaussian, then we have

$$p(Z|I) = \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{\infty} dX e^{-(X-\mu_x)^2/2\sigma_x^2} \times e^{-(X-Z-\mu_y)^2/2\sigma_y^2}$$

One can do some pretty boring algebra at this point (factoring the exponents), or use a program like `xmaxima`:

(C1) `ASSUME_POS:TRUE;`

(D1) `TRUE`

(C2) `1/(2*%PI)/sx/sy*integrate(exp(-(x-xo)^2/(2*sx^2))*  
exp(-(x-z-yo)^2/(2*sy^2)),x,-inf,inf);`

$$-\frac{z^2 + (2y_0 - 2x_0)z + y_0^2 - 2x_0y_0 + x_0^2}{2sy^2 + 2sx^2}$$

(D2) `SQRT(2) %E`  
-----  
`2 SQRT(%PI) SQRT(sy^2 + sx^2)`

(C3) `factor(z^2+(2*yo-2*xo)*z+yo^2-2*xo*yo+xo^2);`

(D3) `(z + yo - xo)^2`

So we get

$$\begin{aligned} p(Z|I) &= \frac{1}{\sqrt{2\pi(\sigma_x^2 + \sigma_y^2)}} e^{-(z-(\mu_x-\mu_y))^2/2(\sigma_x^2+\sigma_y^2)} \\ &= \frac{1}{\sqrt{2\pi\sigma_z^2}} e^{-(z-\mu_z)^2/2\sigma_z^2} \end{aligned}$$

where  $\mu_z \equiv \mu_x - \mu_y$   
 $\sigma_z^2 \equiv \sigma_x^2 + \sigma_y^2$

### C.1.2 Continuing with Paired Data

Changing variables to  $\delta_k$ , it is clear that the likelihood for  $\delta_k$  is the same form as  $\delta_x$  and  $\delta_y$ . Thus we have the *exact same* results on the paired difference, both for known and unknown  $\sigma$ , quoted in Section B.1 and B.2.

## C.2 Difference of Means, $\delta \equiv \mu_x - \mu_y$ , known $\sigma_x$ and $\sigma_y$

Again, the change of variables trick works, but since we are given the means ( $\mu_x$  and  $\mu_y$ ) we need to use the posterior distributions,  $p(\mu_x|\mathbf{x}, \sigma_x, I)$  and  $p(\mu_y|\mathbf{y}, \sigma_y, I)$ .

$$p(\mu_x | \mathbf{x}, \sigma_x, I) = \sqrt{\frac{n}{2\pi\sigma_x^2}} e^{-n(\bar{x} - \mu_x)^2 / 2\sigma_x^2}$$

$$p(\mu_y | \mathbf{y}, \sigma_y, I) = \sqrt{\frac{m}{2\pi\sigma_y^2}} e^{-n(\bar{y} - \mu_y)^2 / 2\sigma_y^2}$$

Performing the change of variables to  $\delta \equiv \mu_x - \mu_y$  we get

$$p(\delta | \mathbf{x}, \mathbf{y}, \sigma_x, \sigma_y, I) = \frac{\sqrt{nm}}{2\pi\sigma_x\sigma_y} \int d\mu_y e^{-n(\bar{x} - \delta - \mu_y)^2 / 2\sigma_x^2} e^{-m(\bar{y} - \mu_y)^2 / 2\sigma_y^2}$$

Again, using `xmaxima`

```
(C1) ASSUME_POS:TRUE;
```

```
(D1)                                     TRUE
```

```
(C2) f(d):=sqrt(n*m)/(2*%PI*sx*sy)*integrate(exp(-n*(xbar-d-my)^2/(2*sx^2))*
      exp(-m*(ybar-my)^2/(2*sy^2)),my,-inf,inf);
```

```
f(d);
```

```
(D2) f(d) := ----- INTEGRATE(EXP(-----)
      Sqrt(n m)          (- n) (xbar - d - my)
      2 %PI sx sy          2
                          2
                          2 sx
                          2
                          (- m) (ybar - my)
                          EXP(-----), my, - INF, INF)
                          2
                          2 sy
```

```
(C3)
```

```
(D3) Sqrt(2) Sqrt(m) Sqrt(n) EXPT(%E, - (m n ybar
      2
      + (- 2 m n xbar + 2 d m n) ybar + m n xbar  - 2 d m n xbar + d  m n)
      2
      / (2 n sy  + 2 m sx )) / (2 Sqrt(%PI) Sqrt(n sy  + m sx ))
(C4) factor((m*n)*ybar^2+(-2*m*n*xbar+2*d*m*n)*ybar+m*n*xbar^2-2*d*m*n*xbar+m*n*d^2);
```

```
(D4)                                     2
      m n (ybar - xbar + d)
```

Rewritten, this is

$$p(\delta|\mathbf{x}, \mathbf{y}, \sigma_x, \sigma_y, I) = \sqrt{\frac{nm}{2\pi(n\sigma_x^2 + m\sigma_y^2)}} e^{-mn(\delta - (\bar{x} - \bar{y}))^2 / 2(n\sigma_x^2 + m\sigma_y^2)}$$

or

$$\begin{aligned} \mu_\delta &\equiv \mu_x - \mu_y \\ \sigma_\delta &\equiv \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m} \\ p(\delta|\mathbf{x}, \mathbf{y}, \sigma_x, \sigma_y, I) &= \frac{1}{\sqrt{2\pi\sigma_\delta^2}} e^{-(\delta - \mu_\delta)^2 / 2\sigma_\delta^2} \end{aligned}$$

### C.3 Difference of Means, $\delta \equiv \mu_x - \mu_y$ , unknown $\sigma_x$ and $\sigma_y$

Making definitions as before for the  $t$  distribution for each variable

$$\begin{aligned} t_x &\equiv \frac{\mu_x - \bar{x}}{S_x / \sqrt{n}} \\ t_y &\equiv \frac{\mu_y - \bar{y}}{S_y / \sqrt{n}} \\ S_x^2 &\equiv \frac{1}{(n-1)} \sum_{k=1}^n (x_k - \mu_x)^2 \\ S_y^2 &\equiv \frac{1}{(m-1)} \sum_{k=1}^m (y_k - \mu_y)^2 \end{aligned}$$

From the addition of variables we get

$$\begin{aligned} t &\equiv \frac{\delta - (\bar{x} - \bar{y})}{\sqrt{S_x^2/m + S_y^2/n}} \\ \tan \theta &\equiv \frac{S_x / \sqrt{n}}{S_y / \sqrt{m}} \end{aligned}$$

$\tan \theta$  depends on the data, and  $t_x$ , and  $t_y$  are known, so the distribution for  $t$  should be known. It is named the Behren's distribution.

### C.4 Ratio of Two Variances $\kappa \equiv \sigma_x^2 / \sigma_y^2$

From Section B.2.4 we have

$$p(\sigma_x|\mathbf{x}, I) \propto \frac{1}{\sigma_x^n} e^{-V_x/2\sigma_x^2}$$

For independent  $\mathbf{x}$  and  $\mathbf{y}$  we have

$$p(\sigma_x, \sigma_y | \mathbf{x}, \mathbf{y}, I) \propto \frac{1}{\sigma_x^n} \frac{1}{\sigma_y^m} e^{-V_x/2\sigma_x^2} e^{-V_y/2\sigma_y^2}$$

Changing variables to  $\kappa \equiv \sigma_x^2/\sigma_y^2$ , we have the following definitions

$$\begin{aligned} \kappa &\equiv \sigma_x^2/\sigma_y^2 \\ \sigma_x &= \sigma_y \kappa^{1/2} \\ \sigma_y &= \sigma_x \kappa^{-1/2} \end{aligned}$$

and then transform the posterior

$$\begin{aligned} p(\kappa, \sigma_x | \mathbf{x}, \mathbf{y}, I) &= p(\sigma_x, \sigma_y | \mathbf{x}, \mathbf{y}, I) \times \left| \frac{\partial(\sigma_x, \sigma_y)}{\partial(\kappa, \sigma_x)} \right| \\ &= p(\sigma_x, \sigma_y | \mathbf{x}, \mathbf{y}, I) \times \left| \begin{array}{cc} \frac{\partial \sigma_x}{\partial \kappa} & \frac{\partial \sigma_x}{\partial \sigma_x} \\ \frac{\partial \sigma_y}{\partial \kappa} & \frac{\partial \sigma_y}{\partial \sigma_x} \end{array} \right| \\ &= p(\sigma_x, \sigma_y | \mathbf{x}, \mathbf{y}, I) \times \left| \begin{array}{cc} \frac{1}{2} \sigma_y \kappa^{-1/2} & 1 \\ -\frac{1}{2} \sigma_x \kappa^{-3/2} & 0 \end{array} \right| \\ &\propto \frac{1}{\sigma_x^n} \frac{\kappa^{m/2}}{\sigma_x^m} e^{-V_x/2\sigma_x^2} e^{-V_y \kappa / 2\sigma_x^2} \sigma_x \kappa^{-3/2} \end{aligned}$$

Now we integrate out the nuisance parameter,  $\sigma_x$ , to get

$$\begin{aligned} p(\kappa | \mathbf{x}, \mathbf{y}, I) &= \int d\sigma_x p(\kappa, \sigma_x | \mathbf{x}, \mathbf{y}, I) \\ &\propto \kappa^{(m-3)/2} \int d\sigma_x \frac{1}{\sigma_x^{n+m-1}} e^{-(V_x + V_y \kappa)/2\sigma_x^2} \end{aligned}$$

from the integral trick (Appendix E.4) we get

$$p(\kappa | \mathbf{x}, \mathbf{y}, I) \propto \kappa^{(m-3)/2} (V_x + V_y \kappa)^{(n+m-2)/2}$$

A more common form is found with the substitutions

$$\begin{aligned} \eta &\equiv \kappa \times \frac{(V_y/f_y)}{(V_x/f_x)} \\ f_x &\equiv n - 1 \\ f_y &\equiv m - 1 \end{aligned}$$

from which it follows

$$p(\eta | \mathbf{x}, \mathbf{y}, I) \propto \eta^{\frac{f_y}{2}-1} (f_x + f_y \eta)^{(f_x + f_y)/2}$$

which is the commonly used F distribution.

## C.5 Simple Linear Regression, $y_k = mx_k + b + \epsilon$

Given

$$y_k = mx_k + b + \epsilon$$

where the (known) noise term is

$$p(\epsilon|I) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\epsilon/2\sigma^2}$$

then we have

$$\begin{aligned} p(m, b|\mathbf{y}, I) &= \int_0^\infty p(m, b, \sigma|\mathbf{y}, I) d\sigma \\ &= \int_0^\infty p(\mathbf{y}|m, b, \sigma, I) p(m, b, \sigma|I) d\sigma \\ p(y_k|m, b, \sigma, I) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(mx_k + b - y_k)/2\sigma^2} \\ p(\mathbf{y}|m, b, \sigma, I) &\propto \frac{1}{\sigma^N} e^{-\sum (mx_k + b - y_k)^2/2\sigma^2} \end{aligned}$$

As before, with uniform priors, we get (assuming we know  $\sigma$ )

$$\begin{aligned} p(m, b|\mathbf{y}, I) &\propto \frac{1}{\sigma^N} e^{-\sum (mx_k + b - y_k)^2/2\sigma^2} \\ L &= \text{constant} - \sum (mx_k + b - y_k)^2/2\sigma^2 \\ \nabla_{m,b} L &= 0 \text{ (maximum } p = \text{maximum } L = \text{minimum squares)} \end{aligned}$$

Gives the two equations

$$\begin{aligned} \sum (mx_k + b - y_k)x_k &= 0 \\ \sum (mx_k + b - y_k) &= 0 \end{aligned}$$

Define

$$\begin{aligned} v &= \sum x_k^2 \\ c &= \sum x_k y_k \\ \bar{x} &= \frac{1}{N} \sum x_k \\ \bar{y} &= \frac{1}{N} \sum y_k \end{aligned}$$

and we have

$$\begin{aligned} \sum (mx_k + b - y_k)x_k &= 0 \\ vm + N\bar{x}b - c &= 0 \end{aligned}$$

and

$$\begin{aligned}\sum(mx_k + b - y_k) &= 0 \\ N\bar{x}m + Nb - N\bar{y} &= 0\end{aligned}$$

### C.5.1 Quick recipe for solving $2 \times 2$ equations

A nice little trick I learned in high school for quickly solving  $2 \times 2$  equations is to use the determinant. It can be used for any size, but it is particularly expedient for  $2 \times 2$  equations.

1. Write the equations in the following form:

$$\begin{aligned}ax + by &= c \\ dx + ey &= f\end{aligned}$$

2. Form the determinant of the left-hand side parameters like

$$\begin{aligned}D &\equiv \begin{vmatrix} a & b \\ d & e \end{vmatrix} \\ &= ae - bd\end{aligned}$$

3. The solutions are formed by the following ratios

$$\begin{aligned}x &= \frac{\begin{vmatrix} c & b \\ f & e \end{vmatrix}}{D} \\ &= \frac{ce - bf}{ae - bd}\end{aligned}$$

and

$$\begin{aligned}y &= \frac{\begin{vmatrix} a & c \\ d & f \end{vmatrix}}{D} \\ &= \frac{af - cd}{ae - bd}\end{aligned}$$

Notice that the numerators are made in the same way as  $D$ , except that the relevant column (1st column for  $x$ , 2nd for  $y$ ) is replaced with the right-hand side parameters.

Why is this any better than solving for one, and plugging in? I find that the arithmetic in this recipe to be more straightforward, and less prone to careless errors.

### C.5.2 Solution to the Simple Least Squares Problem

So we have

$$\begin{aligned}vm + N\bar{x}b &= c \\ \bar{x}m + b &= \bar{y}\end{aligned}$$

Solving we get

$$\begin{aligned}D &\equiv \begin{vmatrix} v & N\bar{x} \\ \bar{x} & 1 \end{vmatrix} \\ &= v - N(\bar{x})^2 \\ m &= \frac{\begin{vmatrix} c & N\bar{x} \\ \bar{y} & 1 \end{vmatrix}}{D} = \frac{c - N\bar{x}\bar{y}}{v - N(\bar{x})^2} \\ b &= \frac{\begin{vmatrix} v & c \\ \bar{x} & \bar{y} \end{vmatrix}}{D} = \frac{v\bar{y} - c\bar{x}}{v - N(\bar{x})^2}\end{aligned}$$

# Appendix D

## Data

### D.1 Bank Waiting Times (in Minutes)

WaitTime	9.8	2.2	10.7
1.6	2.9	5.8	7.5
6.6	10.9	8.6	7.0
5.6	4.7	2.0	5.3
5.1	3.9	8.0	4.5
3.9	3.2	9.5	4.5
4.6	6.5	1.3	1.8
6.5	2.3	4.8	3.7
6.4	8.7	9.9	4.0
8.6	4.7	7.9	5.7
4.2	3.6	1.1	2.4
5.8	2.7	7.2	9.2
3.4	3.7	2.9	5.0
9.3	0.4	9.1	7.2
7.4	10.2	7.7	7.3
1.8	2.8	4.0	4.1
6.2	11.6	1.4	6.7
5.4	4.3	7.8	3.5
4.9	3.1	5.2	6.3
5.4	5.8	8.4	4.3
0.8	5.6	6.3	3.8
2.5	4.4	4.4	5.1
8.3	4.5	5.2	5.5
6.1	6.7	7.4	
6.3	8.1	6.1	
6.8	4.3	3.8	

### D.2 Customer Satisfaction (7-pt Likert Scale $\times$ 7 responses)

---

Ratings	42	46	48
39	46	43	44
45	40	47	41
38	47	43	45
42	44	41	44
42	43	40	44
41	45	43	44
38	45	44	46
42	40	41	39
46	46	38	41
44	41	43	44
40	43	36	42
39	39	44	47
40	43	44	43
42	46	45	45
45	45	44	
44	45	46	

# Appendix E

## Probability Distributions and Integrals

### E.1 Binomial

The binomial probability distribution is given by

$$P_N(k) = \binom{N}{k} p^k q^{N-k}$$

where  $q = (1 - p)$ . The name comes from the expansion of the binomial  $(p + q)^N$ :

$$(p + q)^N = \sum_{k=0}^N \binom{N}{k} p^k q^{N-k}$$

#### E.1.1 Normalization

We want to verify the normalization condition

$$\sum_{k=0}^N P_N(k) = 1$$

We do this by equating this sum with the binomial expansion

$$\begin{aligned} \sum_{k=0}^N P_N(k) &= \sum_{k=0}^N \binom{N}{k} p^k q^{N-k} = (p + q)^N \\ &= (p + (1 - p))^N = 1 \quad \checkmark \end{aligned} \tag{E.1.1}$$

#### E.1.2 Mean

The mean of the distribution is simply

$$\langle k \rangle = \sum_{k=0}^N k \cdot P_N(k) = \sum_{k=0}^N \binom{N}{k} k p^k q^{N-k}$$

We can perform a trick to relate this sum to the normal binomial sum (Equation E.1.2), by taking a derivative of the sum with respect to  $p$ .

$$kp^k = p \frac{\partial}{\partial p} (p^k)$$

$$p \frac{\partial}{\partial p} \left[ \sum_{k=0}^N \binom{N}{k} p^k q^{N-k} \right] = \sum_{k=0}^N \binom{N}{k} kp^k q^{N-k} = \langle k \rangle$$

Thus we have

$$\begin{aligned} \langle k \rangle &= p \frac{\partial}{\partial p} (p+q)^N \\ &= Np(p+q)^{N-1} = Np(p+(1-p))^{N-1} \\ \langle k \rangle &= Np \quad \checkmark \end{aligned} \tag{E.1.2}$$

### E.1.3 Variance

A similar trick can be performed to get the variance.

$$k^2 p^k = \left( p \frac{\partial}{\partial p} \right)^2 (p^k)$$

Thus we have

$$\begin{aligned} \langle k^2 \rangle &= \left( p \frac{\partial}{\partial p} \right)^2 (p+q)^N \\ &= p(N(p+q)^{N-1} + pN(N-1)(p+q)^{N-2}) \\ &= (Np)^2 + Npq \\ (\delta k)^2 &= \langle k^2 \rangle - \langle k \rangle^2 = Npq \end{aligned} \tag{E.1.3}$$

### E.1.4 Gaussian Approximation

$$\begin{aligned} P_N(k) &= \binom{N}{k} p^k q^{N-k} \\ \mu &\approx Np \\ \sigma^2 &\approx Npq \\ &\approx \frac{1}{\sqrt{2\pi Npq}} e^{-(k-Np)^2/2Npq} \end{aligned}$$

## E.2 Negative Binomial

This distribution is used when you are sampling from a Bournoulli process, but wait until you have a certain number of successes. Waiting for  $h$  successes, you sample  $N$  times. The distribution is

$$P(N) = \binom{N-1}{h-1} p^h (1-p)^{N-h}$$

It is as if you have a binomial situation with  $h-1$  successes, then (knowing that you are going to stop at the last success) multiply by the probability of one more success.

## E.3 Beta

## E.4 Gaussian

The basic Gaussian integral is

$$I = \int_{-\infty}^{+\infty} e^{-ax^2} dx \tag{E.4.4}$$

In order to solve this, let's first consider  $I^2$  which is

$$I^2 = \int_{-\infty}^{+\infty} e^{-ax^2} dx \int_{-\infty}^{+\infty} e^{-ay^2} dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-a(x^2+y^2)} dx dy \tag{E.4.5}$$

The reason for introducing another integration variable ( $y$ ) can be understood if one thinks of the integral as a sum, so the product of two sums is the sum of all of the different products of the individual elements in the sum. Thus the double integral. We now notice that if we were to work in polar coordinates,  $x^2 + y^2$  would become  $r^2$  and the infinitesimal area  $dx dy$  would become  $r dr d\phi$ , and our integral would take on a much easier form

$$\begin{aligned} u &\equiv ar^2 \\ du = 2ardr &\Rightarrow \frac{du}{2a} = r dr \\ I^2 &= \int_0^{\infty} e^{-ar^2} r dr \int_0^{2\pi} d\phi \\ &= 2\pi \int_0^{\infty} e^{-u} du \frac{1}{2a} \\ &= \frac{-\pi}{a} e^{-u} \Big|_0^{\infty} \\ &= 0 - \frac{-\pi}{a} \\ &= \frac{\pi}{a} \\ I &= \sqrt{\frac{\pi}{a}} \end{aligned} \tag{E.4.6}$$

The more general form the the Gaussian integral is

$$I = \int_{-\infty}^{+\infty} e^{-(ax^2+bx)} dx \quad (\text{E.4.7})$$

Now we have to beat Equation E.4.7 into the form of Equation E.4.4 so we can use the solution we arrived at earlier. To do this we complete the square in the exponent

$$\begin{aligned} I &= \int_{-\infty}^{+\infty} e^{-(ax^2+bx)} dx \\ &= \int_{-\infty}^{+\infty} e^{-(ax^2+bx+\frac{b^2}{4a}-\frac{b^2}{4a})} dx \\ &= \int_{-\infty}^{+\infty} e^{-(\sqrt{a}x+\frac{b}{2\sqrt{a}})^2+\frac{b^2}{4a}} dx \\ &= e^{\frac{b^2}{4a}} \int_{-\infty}^{+\infty} e^{-u^2} du \frac{1}{\sqrt{a}} \\ (\text{Using Equation E.4.6}) &= e^{\frac{b^2}{4a}} \sqrt{\frac{\pi}{a}} \end{aligned} \quad (\text{E.4.8})$$

### E.4.1 Aside about Cool Tricks for Gaussian Integrals

Integrals of the form of a simple power ( $t^{N-2}$ ) multiplied by a Gaussian:

$$\int_0^{\infty} t^n e^{-at^2} dt \equiv I_n$$

come up frequently.

The basic Gaussian integral is

$$I = \int_{-\infty}^{+\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}} \quad (\text{E.4.9})$$

Now, the power integral we have is from 0 to  $\infty$ , not from  $-\infty$  to  $\infty$ , so the solution is half of this.

Thus,

$$\begin{aligned} I_0 &= \int_0^{\infty} t^0 e^{-at^2} dt \\ &= \frac{1}{2} \sqrt{\frac{\pi}{a}} \end{aligned}$$

The next one is even easier.

$$\begin{aligned} I_1 &= \int_0^{\infty} t^1 e^{-at^2} dt \\ u &\equiv at^2 \\ du &= 2at dt \\ I_1 &= \int_0^{\infty} \frac{1}{2a} e^{-u} du \\ &= -\frac{1}{2a} e^{-u} \Big|_0^{\infty} \\ &= \frac{1}{2a} \end{aligned}$$

Now, here is the cool trick. Say, I take  $I_0$ , and before solving the integral, I differentiate it with respect to  $a$ . I get

$$\begin{aligned}\frac{dI_0}{da} &= \int_0^\infty t^0 \frac{d}{da} e^{-at^2} dt \\ &= - \int_0^\infty t^2 e^{-at^2} dt\end{aligned}$$

which is just  $-I_2$ ! Now, I can do the same thing *after* the solution, and the answer must be the same. Therefore

$$\begin{aligned}I_2 &= -\frac{d}{da} I_0 \\ &= \frac{\sqrt{\pi}}{4} a^{-3/2}\end{aligned}$$

The same trick can be done to get  $I_3$  from  $I_1$

$$\begin{aligned}I_3 &= -\frac{d}{da} I_1 \\ &= \frac{1}{2} a^{-2}\end{aligned}$$

Following this process, it is easy to show that<sup>1</sup>

$$I_n \equiv \int_0^\infty t^n e^{-at^2} dt \quad (\text{E.4.10})$$

$$\propto a^{-(n+1)/2} \quad (\text{E.4.11})$$

$$\quad (\text{E.4.12})$$

Another common integral is

$$J_n \equiv \int_0^\infty \frac{1}{x^n} e^{-a/x^2} dx$$

with the substitution  $t \equiv 1/x$  and  $dt = -dx/x^2$  we get

$$J_n \equiv \int_0^\infty t^{n-2} e^{-at^2} dt$$

From the above integral this becomes

$$J_n \equiv \int_0^\infty \frac{1}{x^n} e^{-a/x^2} dx \quad (\text{E.4.13})$$

$$= a^{-(n-1)/2} \quad (\text{E.4.14})$$

<sup>1</sup>The full solution is  $I_n = a^{-(n+1)/2} \Gamma((n+1)/2)/2$

# Bibliography

Bernoulli (1713). *Ars Conjectandi (The Art of Conjecture)*.

Bowerman and O'Connell (2003). *Business Statistics in Practice*. McGraw-Hill.

Jaynes, E. T. (1957). Information theory and statistical mechanics I. *Phys. Rev.*, 106:620–530.

Jaynes, E. T. (1976). Confidence intervals vs bayesian intervals. In Harper, W. L. and Hooker, C. A., editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*. D. Reidel, Dordrecht. <http://bayes.wustl.edu/etj/articles/confidence.pdf>.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge. Edited by G. Larry Bretthorst.

Jeffreys, H. (1939). *Theory of Probability*. Oxford Univ. Press. 3rd edition reprinted 1985.

Lee, P. M. (1989). *Bayesian Statistics: An Introduction*. Oxford University Press.

Lindley, D. V. and Phillips, L. D. (1976). Inference for a bernoulli process (a bayesian view). *The American Statistician*, 30(3):112–119.

Loredo, T. J. (1990). From Laplace to supernova SN 1987A: Bayesian inference in astrophysics. In Fougere, P., editor, *Maximum Entropy and Bayesian Methods, Dartmouth, U.S.A., 1989*, pages 81–142. Kluwer.

Sivia, D. S. (1996). *Data Analysis. A Bayesian Tutorial*. Oxford University Press.