3a Probability

a. Definitions

We are inundated with probabilities in environmental fields as well as society. The chance of rain is 50%- what does that mean? The chance of lung cancer in bald males who smoke is XX%. Probabilities should be defined carefully. We begin with some definitions.

- Event- set or class or group of possible uncertain outcomes. Rain/no rain. Temperature greater than 50^{o}F, etc.
- Elementary event- cannot be decomposed into other events
- Compound event- decomposable into 2 or more elementary events or other compound events
- Null event- that which cannot occur

Example: roll 6 sided die. (1) elementary event- 1 spot comes up; (2) compound event- odd number of spots comes up (1, 3, or 5); (3) null event- getting a 7 on a 6 sided die.

Will precipitation occur tomorrow? That is an elementary event if the only other choice is no precipitation. However, a compound event would be: will precipitation greater than 0.1 inch occur (it could rain more or could rain less or not at all) or will it snow or rain or both?

- S- Sample or event space. Set of all possible elementary events or the largest possible compound event
- Mutually exclusive- two events that cannot occur at the same time
- Mutually exclusive and collectively exhaustive events (MECE)- no more than 1 event can occur and at least one event will occur

b. Venn diagrams

Venn diagrams are a convenient way to display the sample space and make sense of the event outcomes that are possible. The NCDC storm event climatology (http://www.ncdc.noaa.gov/stormevents/) is a rich resource for examining weather events. I went through the reports for Salt Lake County from the NCDC Storm Event climatology and counted up the number of cases reported of winter and summer (convective) storms and those storms with property damage greater than $5000 during the thirteen year period 1993-2005. Now, I made some assumptions along the way as far as how to count events- some winter storms events may have been multiple day events, for example, and I counted lightning as being associated with convective storms. I ignored some iffy cases where it could have been a convective winter storm. Property damage has occurred from “other” storms and obviously the results might have been different if I used another $ damage threshold. In any event, there were a total of 142 winter storms and 83 summer storms as defined by me. 79 winter storms had damage in excess of $5000 while 25 summer storms had damage of similar amount. Given the nearly 5000 days during the 13 year record, these major weather events as defined by NCDC are not very common in Salt Lake County. The Venn diagram helps to highlight that winter storms are associated with property damage more frequently than summer storms in Salt Lake County and, as defined here, winter and summer storms are obviously mutually exclusive.

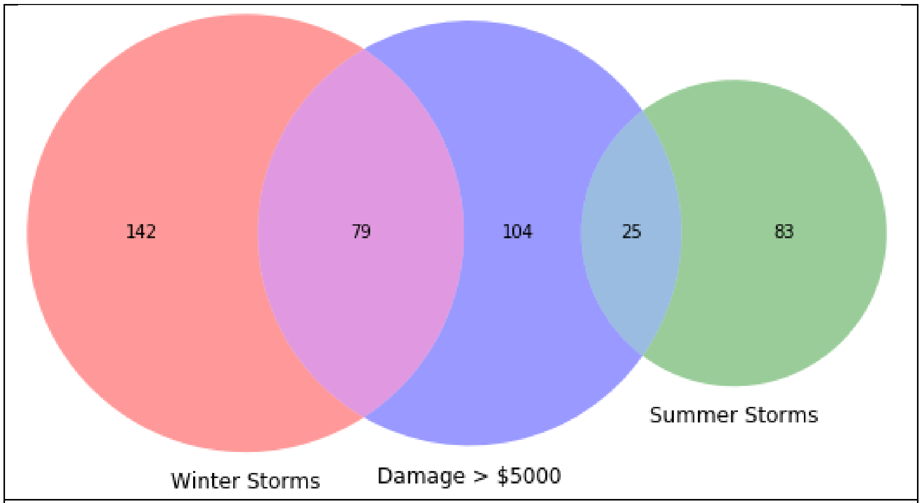
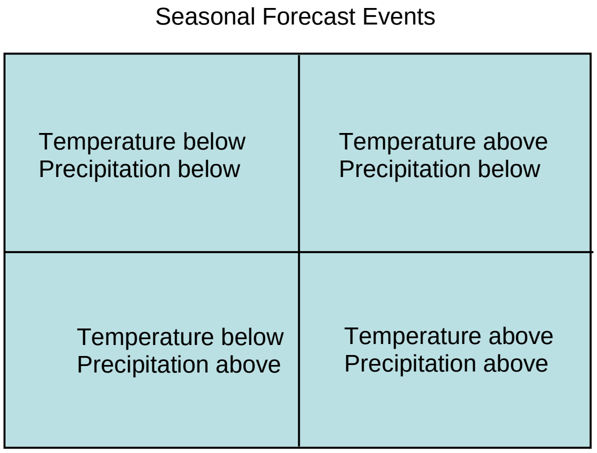


Figure 3.1. Venn diagram of major storms in Salt Lake County during 1993-2005.

Venn diagrams are useful for categorizing events that fall into clear categories and they don't need to be done in terms of circles. Consider Fig 3.2 that shows the possible MECE for a seasonal forecast of above/below normal temperature and precipitation for a specific location. All four possibilities are shown and the probability of each event will depend on the situation and location.



Seasonal Forecast Events

Temperature below Precipitation below	Temperature above Precipitation below
Temperature below Precipitation above	Temperature above Precipitation above

Figure 3.2. MECE possibilities for seasonal forecasts of temperature and precpitation anomalies for a specific location.

c. Probability Concepts

The following are pretty obvious, but when you get mathematicians involved, they have to have "axioms", lemmas, etc.

- probability of any event is nonnegative. In English: an event has to happen or else it is not an event
- probability of the compound event S is 1 or 100%. The probability that all events will happen is 1.
- probability that one or the other of two mutually exclusive events is the sum of their individual probabilities.

Definitions:

- Let E- event
- Pr{E}- probability of Event E; $0 \le \Pr\{E\} \le 1$
- Pr{E}=0 event does not occur
- Pr{E}=1 absolutely sure that event will occur

There are two approaches to probabilities: the frequency view and the Bayesian view. Which approach is used depends on the type of problem being investigated.

Frequency view- probability of an event is its relative frequency after many, many trials

- a- number of occurrences of E
- n- number of opportunities for E to take place
- a/n – relative frequency of event E occurring
- Pr{E} $\rightarrow$ a/n as n $\rightarrow \infty$
- Or a = outcomes = n Pr{E}

Examples: role a die. We expect the 6 spot to come up 1/6 times or 1 time every 6 opportunities. If we role the die 100 times, we expect the 6 to come up 16-17 times. What are the odds of drawing an ace? 4/52. So once every 13 times we expect to draw an ace. However, what we expect and what actually happens are clearly different things, that's where chance/randomness comes into play.

Bayesian view- probability represents the degree of belief or quantifiable judgement of a particular individual about an outcome of an uncertain event

- this approach recognizes that some events occur so rarely that there is no long-term probability estimate that are relevant
- Bookies make odds all the time based on their evaluation of the odds of winning for a particular team- it is not based on a large sample
- Two individuals can have different probabilities for same outcome

More concepts

- If event $\{E_2\}$ occurs whenever $\{E_1\}$ occurs, then $\{E_1\}$ is a subset of $\{E_2\}$
- Example: $\{E_1\}$- temperature below freezing; $\{E_2\}$- temperature below 50F, then $\Pr\{E_1\} \leq \Pr\{E_2\}$
- The complement of {E} is that event $\{E\}^c$ that does not occur
- $\Pr\{E\}^c = 1 - \Pr\{E\}$

What is the probability that $\{E_1\}$ and $\{E_2\}$ occur, that is, the intersection between the two events?

- $\Pr\{E_1 \cap E_2\}$ = joint probability that $\{E_1\}$ and $\{E_2\}$ will occur *(3.c.1)*
- $\Pr\{E_1 \cap E_2\} = 0$ if $\{E_1\}$and $\{E_2\}$are mutually exclusive
    - Example: if $\{E_1\}$is the occurrence of temperature below freezing and $\{E_2\}$ is the occurrence of temperature above 50°F, then their joint probability is 0.

Let's return to the Venn diagram of the weather events in Salt Lake County. In the way that I went through the sample, the winter storms and convective storms are mutually exclusive, so there is no overlap between those two events. Assuming, that winter storms occur only during the winter half of the year and that convective storms occur only in the summer half (not great assumptions!), then the number of opportunities is order 180 days x 13 years= 2340 opportunities. Also, remember that some of the winter storms could be multiple day events, so I have some uncertainty and error in my results.

- $\{E_1\}$- occurrence of winter storms = 142
- $\Pr\{E_1\}$= 142/2340 = .061

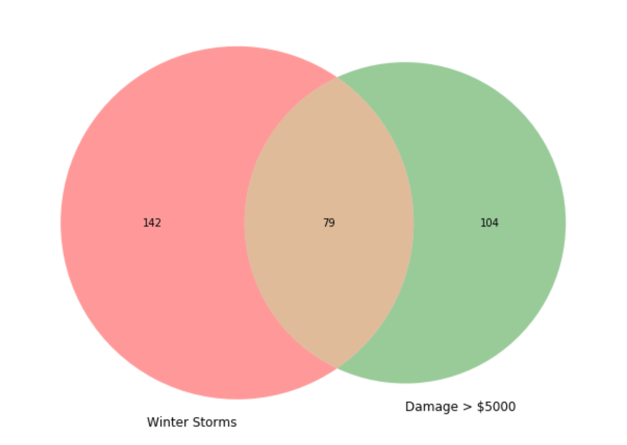


Figure 3.3. The overlap ($\Pr\{E_1 \cap E_2\}$) has to be subtracted from the sum of the probabilities of each event occurring in order to determine if either occur ($\Pr\{E_1 \cup E_2\}$).

•$\Pr\{E_1\}^c = .939$
•$\{E_2\}$- occurrence of summer convective storms = 83
•$\Pr\{E_2\}= 83/2340 = .035$
•$\Pr\{E_2\}^c = .965$
•$\{E_3\}$- occurrence of property damage = 113
•$\Pr\{E_3\}= 113/2340 = .048$
•$\Pr\{E_3\}^c = .952$
•$\Pr\{E_1 \cap E_2\} = 0$
•$\Pr\{E_1 \cap E_3\} = 79/2340 = .034$
•$\Pr\{E_2 \cap E_3\} = 25/2340 = .011$

What is the probability that $\{E_1\}$ OR $\{E_2\}$ will occur? That is, one the other, or both will occur. This is referred to as the "union" of the two events.

- $\Pr\{E_1 \cup E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1 \cap E_2\}$ *(3.c.2)*
- As can be seen visually from the Venn diagram to the right, the joint probability is counted twice when the individual probabilities are summed, so it is subtracted once.

It is worthwhile to see how that is true algebraically as well. Add up each probability separately:

$\Pr\{E_1 \cup E_2\} = \Pr\{E_1\} - \Pr\{E_1 \cap E_2\} + \Pr\{E_1 \cap E_2\} + \Pr\{E_2\} - \Pr\{E_1 \cap E_2\}$ or
$\Pr\{E_1 \cup E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1 \cap E_2\}$

- If $\{E_1\}$ and $\{E_2\}$ are mutually exclusive, then $\Pr\{E_1 \cup E_2\} = \Pr\{E_1\} + \Pr\{E_2\}$

A couple more identities that you should be able to visualize from a Venn diagram:

- $\Pr\{(E_1 \cup E_2)^c\} = \Pr\{ E_1^c \cap E_2^{\,c}\}$
  - that is the area outside of both circles
- $\Pr\{(E_1 \cap E_2)^c\} = \Pr\{ E_1^c \cup E_2^{\,c}\}$
  - this one is a little harder to visualize; it is everything outside of the intersection of the two events

Returning to the Salt Lake County storm data:

- $\Pr\{E_1 \cup E_2\}$ = Pr{winter storms} + Pr{convective storms} - Pr{winter storms ∩ convective storms} = .096 since the two events are mutually exclusive the last term is 0
- $\Pr\{E_1 \cup E_3\} = .061 + .048 - .034 = .082$ = Pr{winter storms or property damage or both}
- $\Pr\{E_2 \cup E_3\} = .035 + .048 - .011 = .072$ = Pr{summer storms or property damage or both}

d. Conditional Probability

The storm data example for Salt Lake County indicates that some winter storms lead to expensive damage. So, given that a winter storm has occurred, what is the probability that damage has occurred? We are now limiting our sample to a smaller number of events, only the 142 winter storms. So, the probability is now the 79 damaging winter storms divided by the 142 total winter storms or 56%.

- Conditional probability: probability that $\{E_2\}$ will occur given that $\{E_1\}$ has occurred
- $\Pr\{E_2 | E_1\} = \Pr\{E_1 \cap E_2\} / \Pr\{E_1\}$ *(3d.1)*

$\{E_1\}$ is called the conditioning event; if it doesn't happen, then we know nothing about the probability that $\{E_2\}$ will happen

Alternatively, we can write:

- $\Pr\{E_1 \cap E_2\} = \Pr\{E_2 | E_1\} \times \Pr\{E_1\} = \Pr\{E_1 | E_2\} \times \Pr\{E_2\}$ *(3d.2)*

Whether we condition from the first or second event to determine the intersection of the two events is our choice and simply depends on the available data.

If two events are completely independent, such that the occurrence of nonoccurrence of one event does not affect the probability of the other, then

- $\Pr\{E_2 | E_1\} = \Pr\{E_2\}$ and $\Pr\{E_1 | E_2\} = \Pr\{E_1\}$

Then,

- $\Pr\{E_1 \cap E_2\} = \Pr\{E_1\} \times \Pr\{E_2\}$ for independent events

If we have a fair coin, then the Pr{head} = .5. The second coin toss does not depend on the first, so Pr{10 heads in a row} = $.5^{10}$

Now for some simple examples using blackjack.

- Pr{ace} = 4/52 ; Pr{10-K} = 16/52 ; Pr{2-9}= 32/52

Deal two cards. What is the probability that you will get twenty-one? The probability for each card are independent events, so Pr{ ace ∩ 10-K} = 4/52 × 16/52 = 2.4%. What is probability of getting twenty-one from 2 successive deals if you reshuffle? = $.024^2$ = .05%. The odds for various combinations are shown below.

First Card

Second card	2-9	10-K	Ace
2-9	.38	.19	.05
10-K	.19	.09	.02
Ace	.05	.02	.01

Every 100 hands, you should get a twenty-one 4 times and about 38 hands with 2 cards being 2-9, etc.

e. Persistence

Persistence is the existence of statistical dependence over time (or space), i.e., that once a phenomenon begins it does not necessarily end before the next observation time or, that the observations at one location are related to the observations at a nearby one. Observations from environmental fields should not be considered to be independent of one another unless care is taken to choose a sample taking into account spatial and temporal dependence.

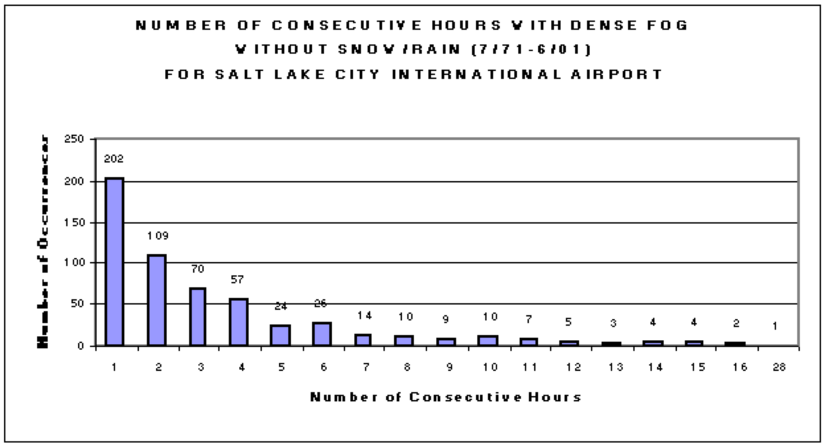


Figure 3.4. Fog climatology at the Salt Lake airport.

Consider the fog climatology shown in Fig. 3.4 that was created for the 2002 Olympics by Jonathan Slemmer. We wanted to provide the Olympic forecasters with some information on the likelihood of persistent heavy fog at the airport. If it happened during the Olympics, then there would have been a bunch of negative consequences with flight delays, etc. (fortunately, it didn't happen during that period). Dense fog doesn't happen very often. The sample here is large: 31 years x 365 days= 11315 days. Dense fog happened over a 2 hour period (Jonathan labels this 1 h of consecutive fog) on only 202 days. So Pr{ 1 hour of consecutive fog} = Pr{1}= 202/11315 = 1.8% of days. If we considered the number of hours in a day as well, then the probability that it would happen in a specific hour would be correspondingly less. Pr{2} = 109/11315 = .96% of days, etc.

The probability that 3 hours in a row of fog (2 consecutive hours) is going to happen is pretty unlikely = .96% for any given day. However, if 1 hour of consecutive fog has already happened, then the odds of it continuing are obviously much higher.

- Pr{2|1} = 109/202 = 53.9%, Pr{3|1} = 70/202 = 35%, Pr{10|5} = 41.6%, etc.

Persistence is a good statistical forecasting baseline. When my family asks me what the weather is going to be like tomorrow, without any other information available, I'm going to say whatever it is today. And, just to let you know, my family doesn't think I'm a very good forecaster. A forecaster adds value when the conditions are present that lead to change and those conditions are correctly recognized as such. While it may be useful to tell an airport operator that the current dense fog has a high likelihood of continuing, more value will be added if the forecaster has information available from which to diagnose when the fog is going to break up.

Later, we will examine ways to estimate the probability of rare events. For example, it is not particularly useful to develop probabilities on the occurrence of dense fog for over 20 consecutive hours based on the 1 event that has happened in our sample. Similarly, we can't wait around for a 100 years to estimate the occurrence of a once in a hundred year flood.

f. Forecast Verification

We'll touch on verification of forecasts from several angles. I'll introduce the difference here between a "measures oriented verification approach" (typically using well-established and often over-used performance metrics such as hit rate or threat score) vs. a "distributions-oriented approach" (where the empirical joint distributions of forecasts and verifying observations are generated).

Let's start with the simple approach that something happens or it doesn't and we forecast it to happen or not. We then count the number of cases for the four possibilities:

		Observed	Observed	**Forecast marginal totals**
		Yes	No	
Forecast	Yes	a	b	**a+b**
Forecast	No	c	d	**c+d**
	Observed marginal totals	**a+c**	**b+d**	**n=a+b+c+d sample size**

First, the marginal totals are important- they are how often something is observed or forecast (or not observed/not forecast). How often do we get the "right" forecasts and how often do we forecast them and they don't happen?

PC = percent correct = $\frac{a+d}{n}$

FAR = false alarm ratio = $\frac{b}{a+b}$

We want the percent correct to be close to 1 and the false alarm ratio to be close to 0. But, what happens when we are verifying categorically forecasts of large precipitation amounts (say over an inch) at Salt Lake City? That doesn't happen very often, but the percent correct may be large because we may be very successful at forecasting when the precipitation is less than an inch (d). Then, to focus on the situations when it either was observed or forecast, the threat score (TS) or critical success index (CSI) is used:

TS = CSI = $\frac{a}{a+b+c}$

The threat score measures the number of correct "yes" forecasts relative to the total number of occasions on which the forecast was forecast or observed. Another metric commonly used is the probability of detection (POD) or hit rate (HR), which identifies how frequently an event is forecast relative to when it is observed:

POD = HR = $\frac{a}{a+c}$

Note when we are using one of the marginal totals in the denominator, we're computing a conditional probability. We can express the hit rate as: given that an event occurs, how often is it correctly forecast? The FAR is: given that an event is forecast, how often did it not happen?

Now let's look at an example using Matt Lammer's research on verifying forecasts made in support of prescribed burn and wildfire operations: http://meso1.chpc.utah.edu/jfsp/

On January 30, 2015, there were 77 forecasts issued in support of prescribed burns nationwide. How often did the forecasters anticipate high wind speeds (≥ 5m/s) later that afternoon to take place relative to what was observed that day?

		Observed	Observed	**Forecast Marginal totals**
		≥ 5m/s	<5 m/s	
Forecast	≥ 5m/s	11	6	**17**
Forecast	<5 m/s	16	44	**60**
	Observed Marginal totals	**27**	**50**	**77**

So, the PC= 71.4%; FAR= 35.3%; TS= 33.3%; and POD = 40.7% . How did they do? Clearly, the percent correct is high because they forecast correctly a lot of cases when the winds were light. The false alarms are not too bad, 6 of the 17 forecasts of high wind. But, the threat score is low because they missed a lot of cases when the winds were observed to be stronger than they were expecting.

Do these forecasts have skill? Statistical skill refers to the relative accuracy of a set of forecasts with respect to reference forecasts (random, persistent, or climatological, for example). The probability of a correct yes forecast by chance (meaning that the observations and forecasts are independent) is just the product of the marginal probabilities of the observations and forecasts:

Random correct yes forecast by chance = $\frac{(a+b)}{n}\frac{(a+c)}{n}$

Random correct no forecast by chance = $\frac{(b+d)}{n}\frac{(c+d)}{n}$

In our case, the odds of having a randomly correct yes forecast is low (7.7%) but the odds of having a randomly correct no forecast is pretty high (50.1%) since it is both observed and forecasted to not be windy frequently.

The most generic of skill scores is of the form: $SS = \frac{(correct\ forecasts - random\ correct\ forecasts)}{(total\ forecasts - random\ correct\ forecasts)}$

The Heidtke Skill Score is of this form and can be computed after some substitutions from the contingency table values as: $HSS = \frac{2(ad-bc)}{(a+c)(b+d)+(a+b)(b+d)}$

In our case, HSS= 31.4%, which is not particularly high and reflects that low wind speed forecasts don't require a lot of skill.

The measures-oriented metrics defined above are ok, but much information is lost by looking only at 2x2 contingency tables. Let's broaden the scope a bit and assume that an accurate forecast is one when the forecast wind speed is within 2 m/s of the observed forecast.
So, most frequently, the forecast errors are up to one m/s weaker than those observed. And, as we determined from the earlier metrics, there is a greater tendency to forecast the wind speeds to be lower than those observed on this particular day. However, we don't know from Fig. 3.4 whether the forecasters do a better job over some ranges of observed wind speeds than others. We can expand the contingency table concept to create a "distributions-oriented" approach to verification as shown in the following table. I've now arbitrarily decided an accurate forecast is within ±2 m/s of that observed and then keep track as well of the sign of the errors that exceed that limit. I've broken up the observed wind speeds into 3 categories as well. Now it is clearer that the forecasters tend to underforecast higher wind speeds and don't ever overforecast high wind speeds, only more moderate ones.

Table 3.1. Distribution-oriented verification of wind forecasts

	E_1	Observed	Observed	Observed	**Error Marginal totals**
E_2		≤3 m/s	3-6 m/s	≥6 m/s	
Error	≤ -2 m/s	0	10	11	**21**
Error	± 2 m/s	22	20	7	**49**
Error	> 2 m/s	0	7	0	**7**
	Observed Marginal totals	**22**	**37**	**18**	**77**

This is a MECE data set for this particular sample of forecasts issued on this single day. If we were to divide the counts in the interior bins by the sample size (77), then those interior bins would be joint probabilities, e.g., 26% of the forecasts were within 2 m/s when the wind speeds were between 3 and 6 m/s (20/77). A lot more information can be gleaned by considering the conditional probabilities as defined by 3c.1 and 3c.2. For example, given that the observed wind speed is greater than 6 m/s ($\Pr\{E_1\} = 18/77 = 23.4\%$), the probability that the forecasters predict the winds to be too light $\Pr\{E_2 | E_1\}$ is:

$$\Pr\{E_2 | E_1\} = \Pr\{E_1 \cap E_2\} / \Pr\{E_1\} = ((11/77)/(18/77)) = 64.7\%$$

And, here's where the interpretation of conditional probabilities can get out of hand. On this particular day, given that the error is greater than +2 m/s, then the probability that the wind is in the 3-6 m/s category is 100%!!

$$\Pr\{E_2 | E_1\} = \Pr\{E_1 \cap E_2\} / \Pr\{E_1\} = ((7/77)/(7/77)) = 100\%$$

Hooray, we can say all forecasters tend to overforecast high winds when the winds are between 3-6 m/s (no- we can't).

Now, imagine only one in ten thousand people will get a particular disease- $\Pr\{E_1\}$. But you hear on the news that 50% of the people that come down with the disease ate jello that day- Pr{E_{2-}- ate jello | E_1}. Should you stop eating jello to avoid catching the disease? $\Pr\{E_1 \cap E_2\} =$

$\Pr\{E_2 \mid E_1\} \times \Pr\{E_1\} = .50 * .0001 = .005\%$ Don’t focus on that eating jello seems to cause an alarming increase in risk; the more important issue is the low risk factor for this particular disease under any circumstance.

g. Summary

Probabilities are at the heart of modern weather forecasting as well as many other environmental applications. While many applications and users will continue to expect to hear on the radio what the temperature will be at 4 PM tomorrow, the underlying information from which a forecaster will base that specific number will likely be probabilistic information. For example, forecasters implicitly use conditional probabilities as part of the forecast preparation. Given the approaching front, and given that a specific model has a known bias in temperature in the prefrontal environment, they expect the temperature to be higher/lower than what would normally take place.