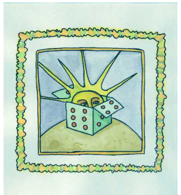
Boulder guide to statistics

Doug Nychka
National Center for Atmospheric Research

- Distributions and densities
- Conditional distributions, Bayes theorem
- Bivariate normal
- Spatial statistics

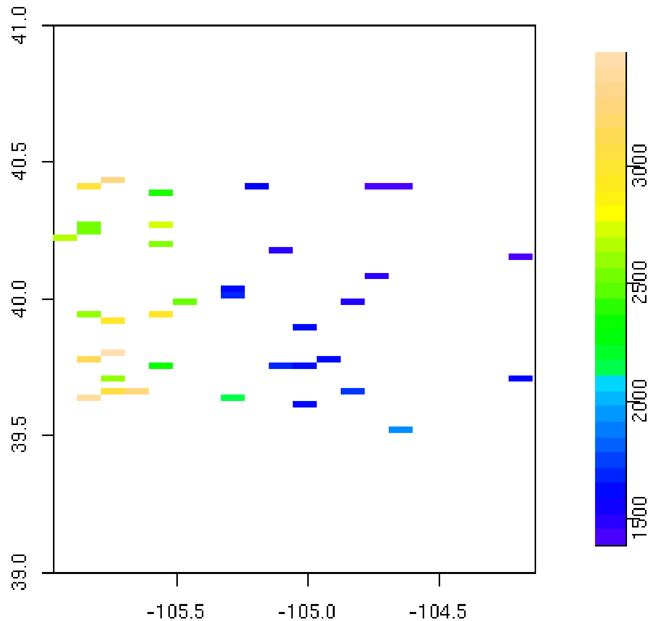
Conditional probability, random sample



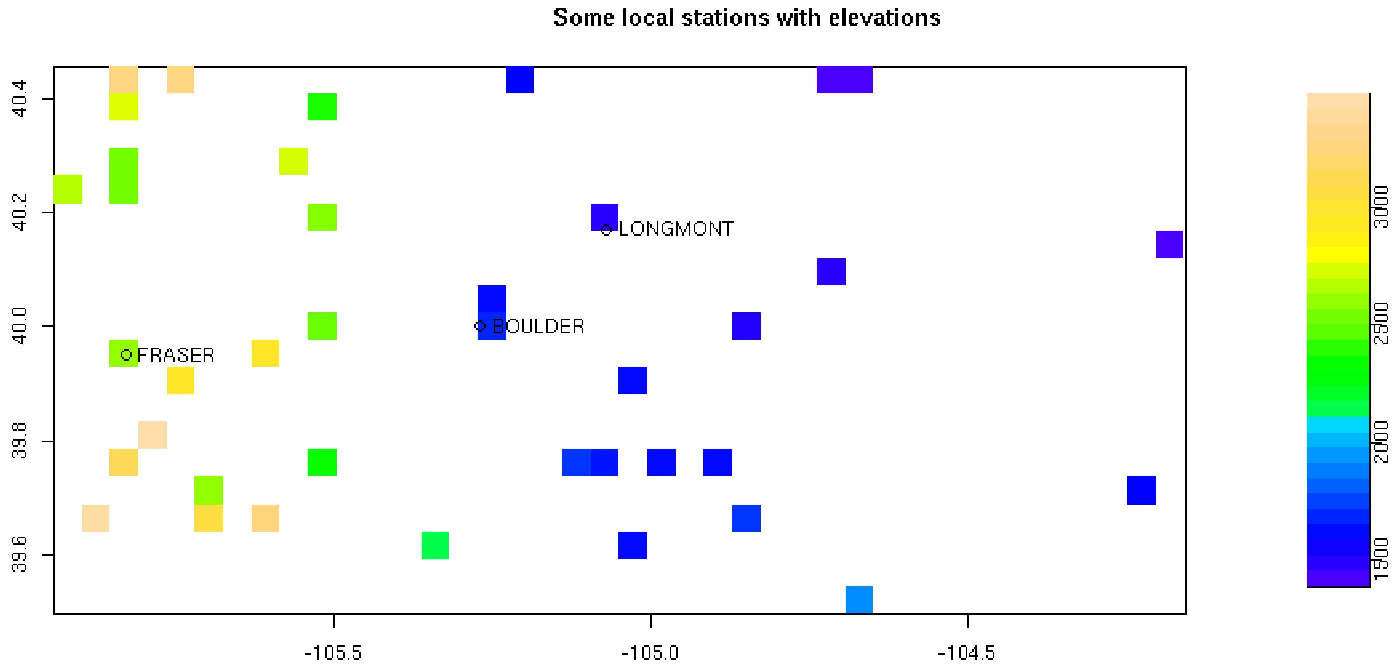
Supported by the National Science Foundation DMS

Overview

As a specific example we will use average July maximum temperatures for an area around Boulder over the period 1895-1997.



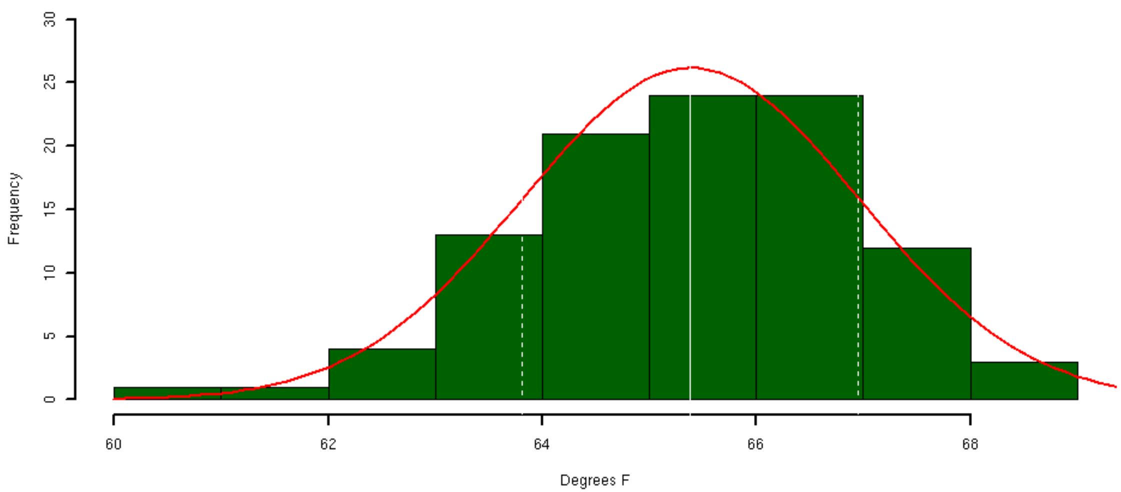
Use the spatial prediction problem to illustrate the concepts of conditional distributions and Bayes theorem.



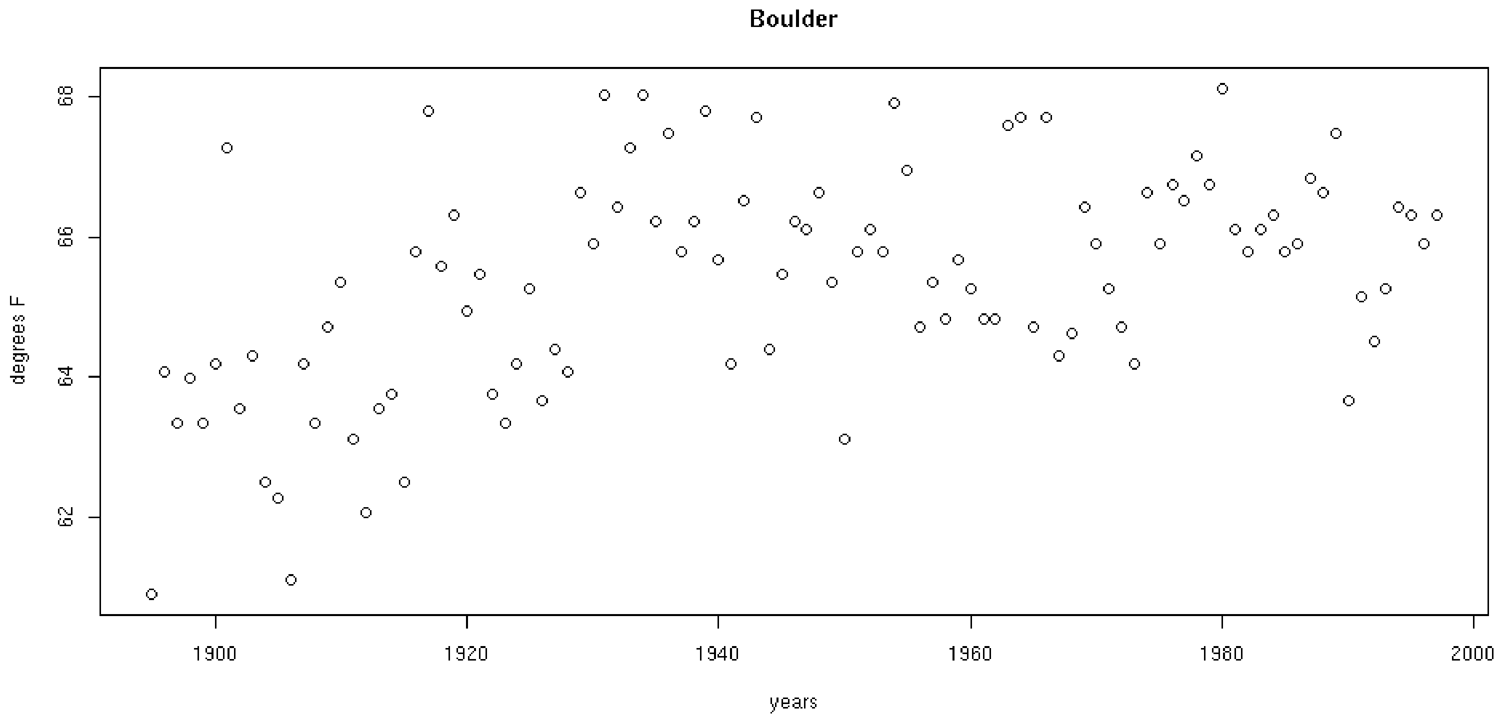
Densities

A probability density function (pdf) is an idealized histogram. It is used to describe probabilities for a random quantity. X = average July temperature for Boulder

$f(x)$ pdf:
Probability that X is in the small interval $[x, x + \Delta]$ is approximately $f(x)\Delta$
Boulder July temps with a normal distribution superimposed:
$(\mu = 65.4, \sigma = 1.6)$



'You can see alot just by looking ...' (Yogi Berra)



I am going to ignore any time trends!

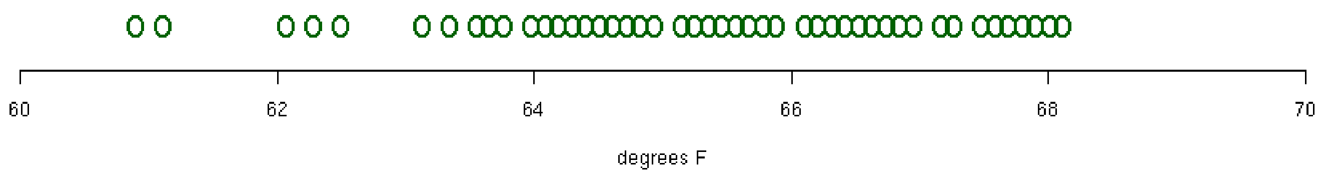
More notes
There are many exotic distributions, *gamma, t, nonparametric, etc.*

Gaussian:

$$f(x) \sim e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

the classic bell-curve shape density, μ and σ are parameters that control the spread and location.

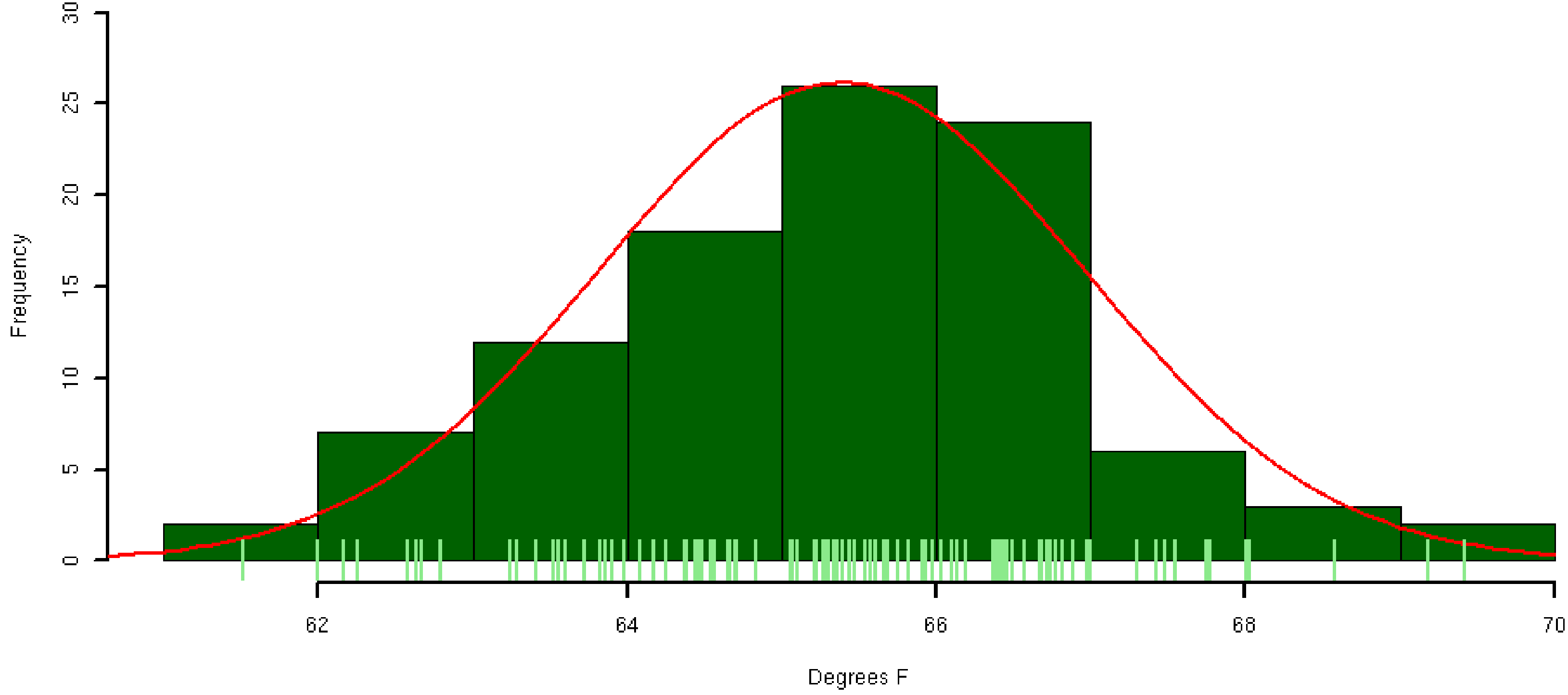
Discrete distribution A finite set of points that are each assigned a probability. Drawing a random sample from a pdf is often a good approximation to the continuous "theoretical" distribution. Here the random sample defines a discrete distribution.



Boulder data (n=103) each point is assigned probability 1/103.

Discrete verses continuous distributions

The continuous normal distribution, a random sample (n=100) drawn from it and the histogram summary.



Statisticians have their moments!

A distribution and a sample both have a *mean* and a *variance* . But they appear to be defined differently and have different interpretations!

Sample mean and variance:

$$\hat{\mu} = \frac{1}{n} \sum_{j=1} n X_j = \sum_{j=1}^{n} X_j (1/n)$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \hat{\mu})^2$$

Mean and variance for a pdf :

$$\mu = \int x f(x) dx$$

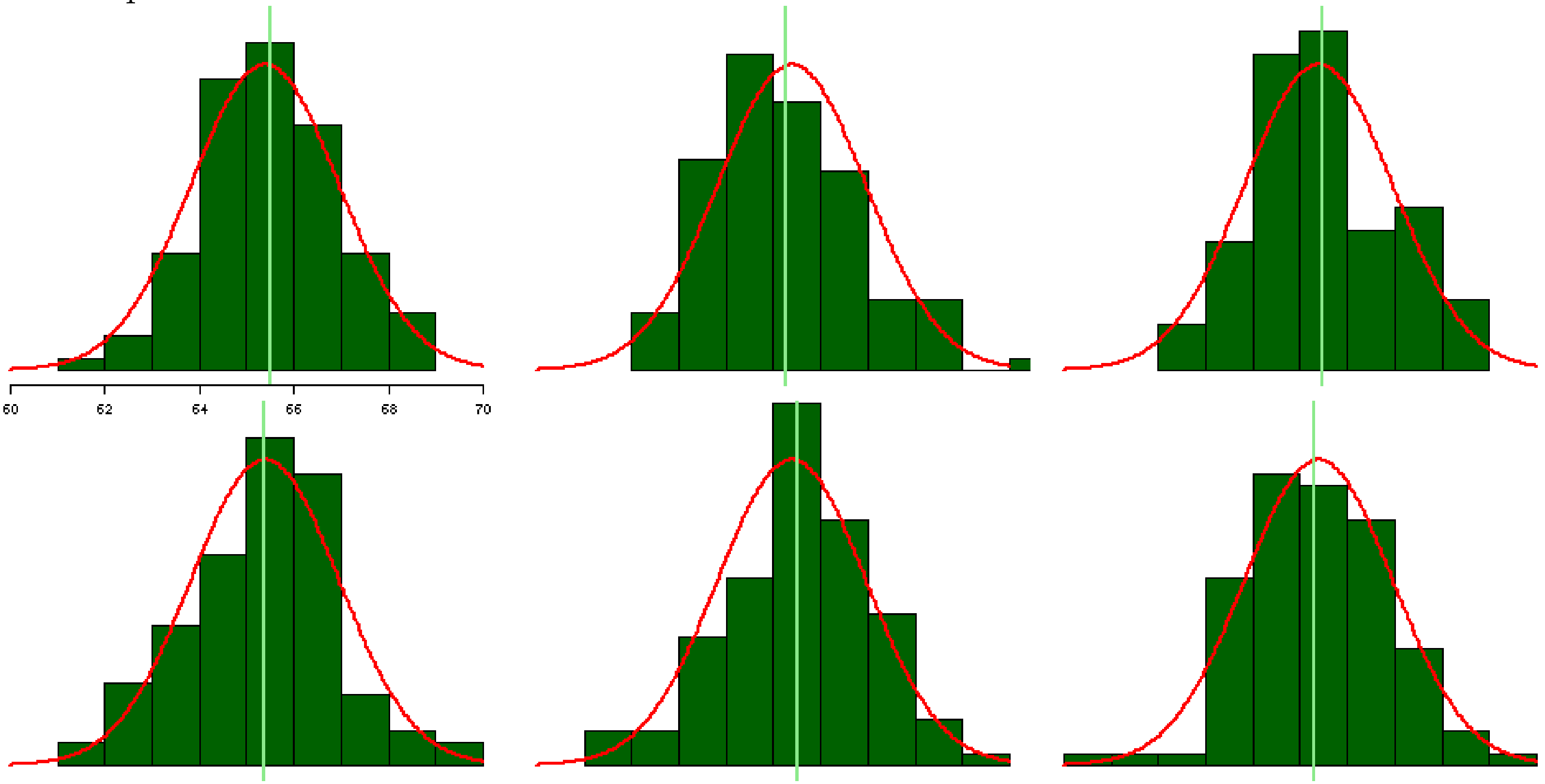
$$\sigma^2 = \int (x-\mu)^2 f(x) dx$$

The connection: If the sample is thought as a discrete distribution where the probability of taking on each data is $1/n$ then the two definitions agree.

The Ensemble Kalman filter uses a discrete distribution at the heart of its statistical algorithm.

Sampling variability

Same thing several times to show the sampling distribution of the histogram and sample mean.



Some other simple remarks:

Mean verses a realization The mean describes the center of the distribution. If X is not known the mean is the best prediction of X in terms of making the error small.
However, the mean not look like a real X value!
e.g. the mean of Boulder July temps (`65.38`) is not equal to any year's value.

Transforming a distribution If X has some pdf and we consider a function of it say $g(x)$ what is the distribution of $g(X)$? e.g. if X is normal then X^2 is χ^2 with 1 degree of freedom.

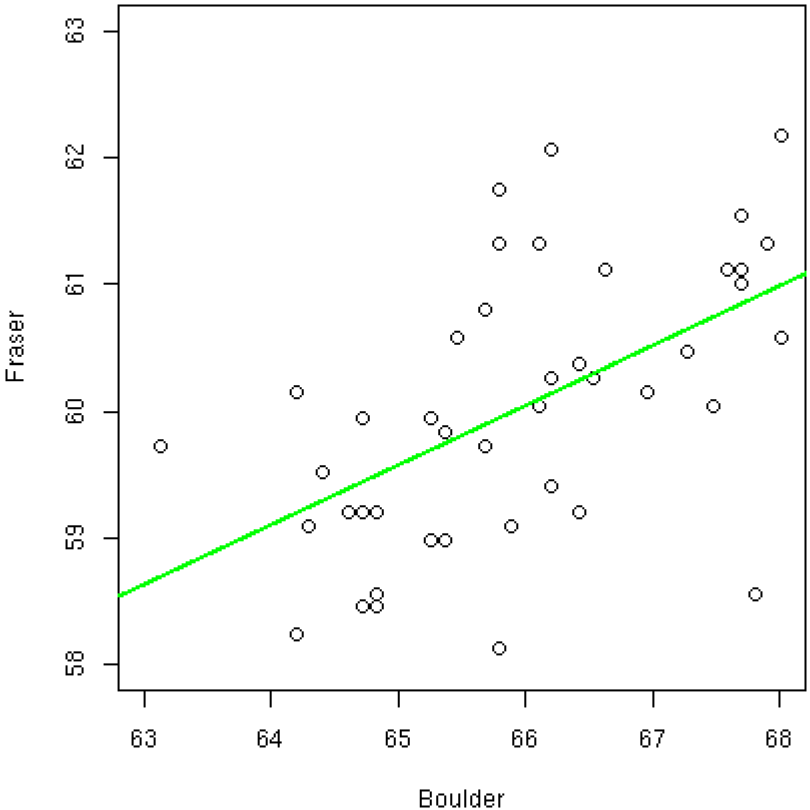
If $X_1, X_2, ..., X_n$ is a random sample from the distribution them $g(X_1), g(X_2), ..., g(X_n)$ is a random sample from the transformed distribution. This is a very useful way to approximate distributions when you need to do a complicated transformation.

For the ensemble Kalman filter g is the forward step of the model, a non-linear function with no closed form.

Multivariate distributions

OK this is really where things get interesting.

A scatterplot of Boulder and Fraser mean July temps



Multivariate distributions

$f(x, y)$ The joint pdf, $f(x, y)$, is defined so that probability of both X and Y being in a small box with sides $[x, x+\Delta]$ and $[y, y+\Delta]$ is approximately $f(x, y)/\Delta^2$.

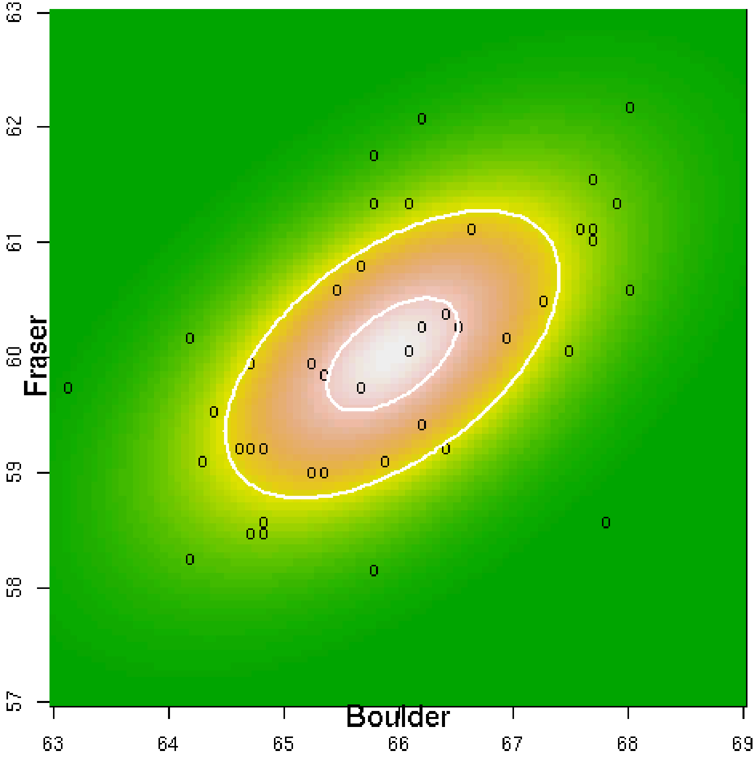
Bivariate normal distribution: Completely described by five parameters: mean(X), mean(Y), VAR(X) , VAR(Y) and COV(X, Y)

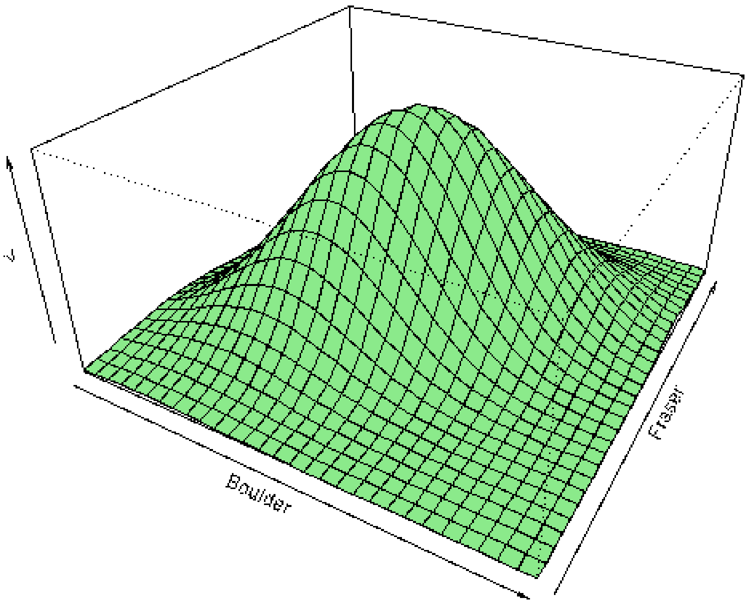
$$\mathrm{COV}(X, Y) = \int (x - \mu_X)(x - \mu_Y) f(x, y) dx dy$$

Covariance matrix: The VARs and COVs are organized in a matrix:

$$\Sigma = \begin{pmatrix} \mathrm{VAR}(X) & \mathrm{COV}(X, Y) \\ \mathrm{COV}(X, Y) & \mathrm{VAR}(Y) \end{pmatrix}$$

Multivariate normal density fit to the Boulder/Fraser data





Conditional distributions

A key step in DA is to determine the distribution of the state of the system given the observed data. The term *given* signals a conditional distribution.

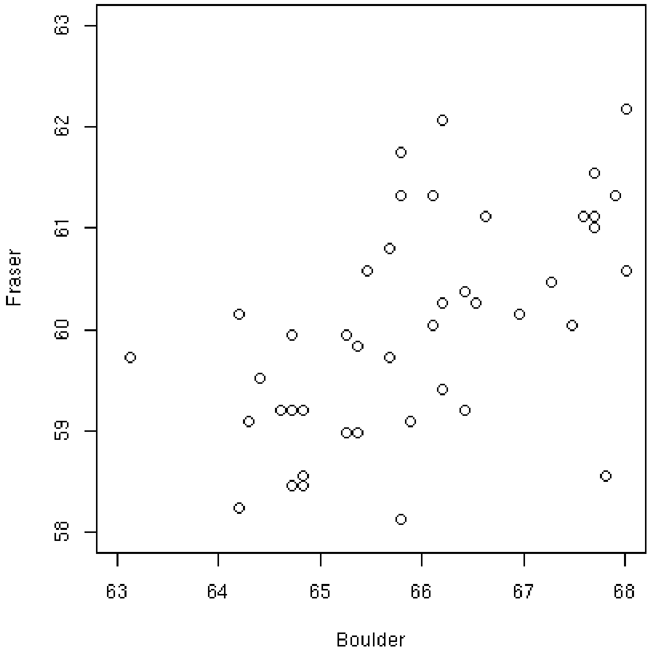
What is the distribution of Fraser temps given that the Boulder temp is `64.5` *or say* `67.5`*?*

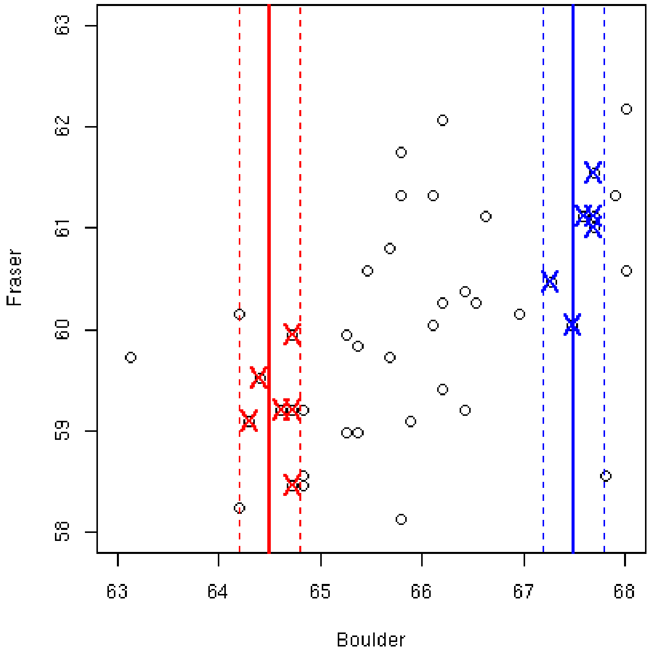
This distribution is different from:

- the joint distribution of both Boulder and Fraser
- the climatological distribution of Fraser (if Fraser and Boulder are not independent).

Motivation using the observed data

Take slices at `65.5` and `68.5`, only consider the data in a neighborhood around each value.

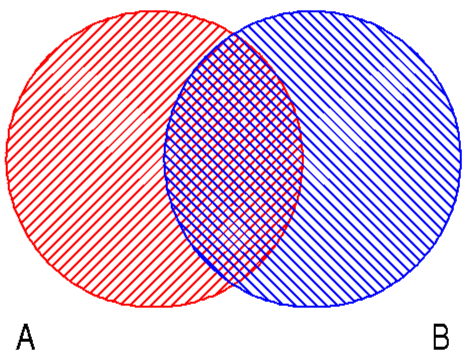




A more formal definition of Conditional Probability

A and B two events
e.g. $A \equiv X \leq 65$, $B \equiv Y \geq 60$

$P(A), P(B)$ denote their probabilities and $P(AB)$ is the probability of both events happening together



Shaded area is $P(AB)$ the conditional probability of B occurring given A occurs is

$$P(B|A) = \frac{P(AB)}{P(A)}$$

The vertical bar is read as *given*.

Conditional densities

$f(x, y)$ the joint pdf for $(X, Y)$and suppose that g(x) is the pdf just for X.

$$f(y|x) = \frac{f(x, y)}{g(x)}$$

Here X is observed (fixed) and we have a distribution for Y.
A useful property of Multivariate normals is that the conditional distributions are also normal.

Some useful notation for pdfs:

- $[Y]$ the pdf for the random variable Y (Fraser temp in this case)
- $[X, Y]$ pdf for joint distribution of X and Y
- $[Y|X]$ conditional pdf for Y *given* X

So the formula for the conditional is:

$$[Y|X] = [X, Y]/[X]$$

Also note that $[X, Y] = [Y|X][X]$

Bayes Theorem

Bayes Theorem gives a way of inverting the conditional information. In bracket notation it is just

$$[Y|X] = \frac{[X|Y][Y]}{[X]}$$

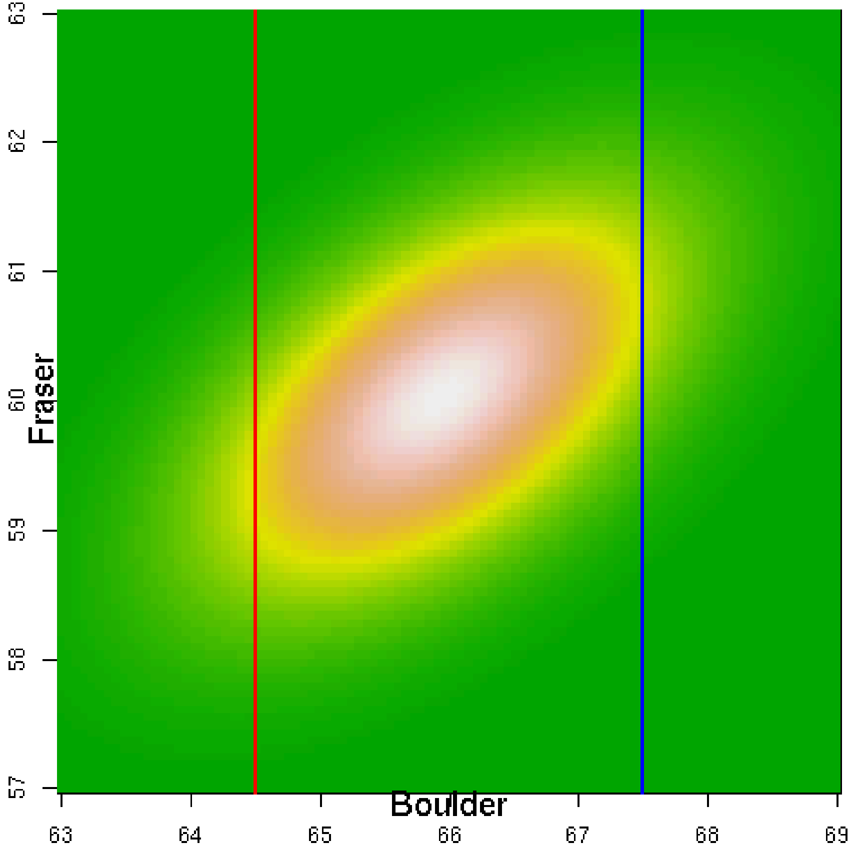
The proof follows by definitions:

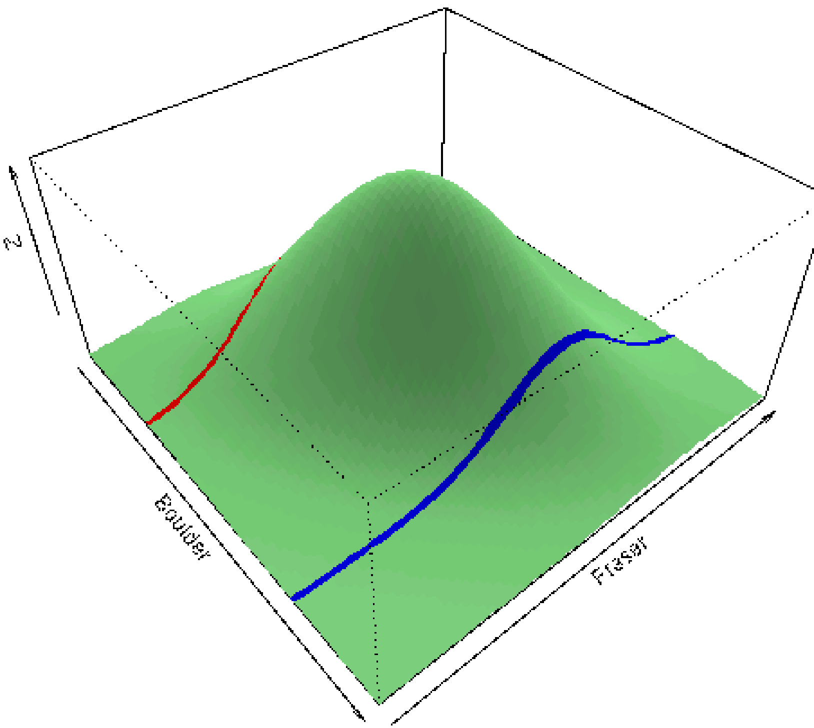
$$[Y|X] = \frac{[X,Y]}{[X]} = \frac{[X|Y][Y]}{[X]}$$

Note that $[Y|X]$ is simply proportional to the joint density where the normalization depends on the values of X. (But in many cases the normalization is difficult to find.)

Conditional densities for the Boulder/Fraser joint pdf

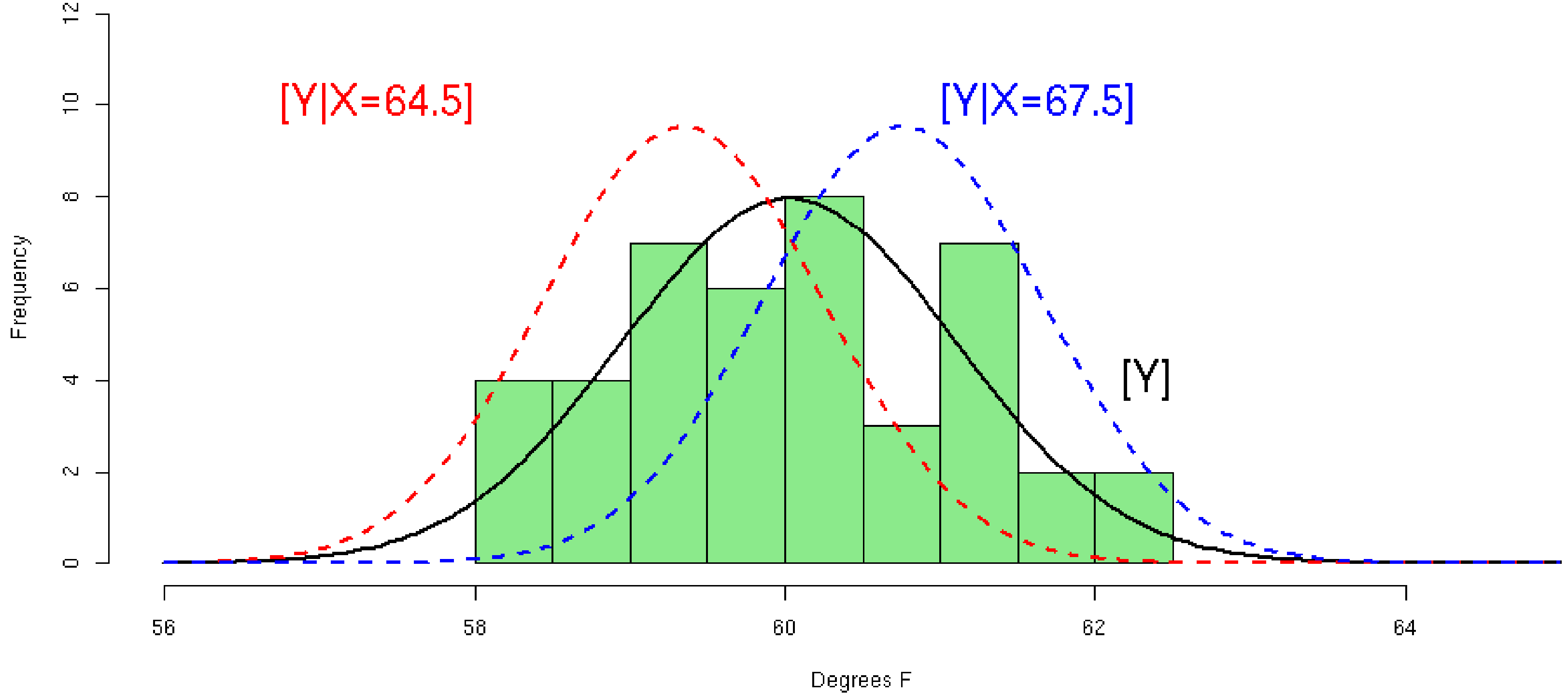
Slicing the surface





Conditional densities for the Boulder/Fraser joint pdf

(Y is Fraser temps and X is Boulder)



Notes on example

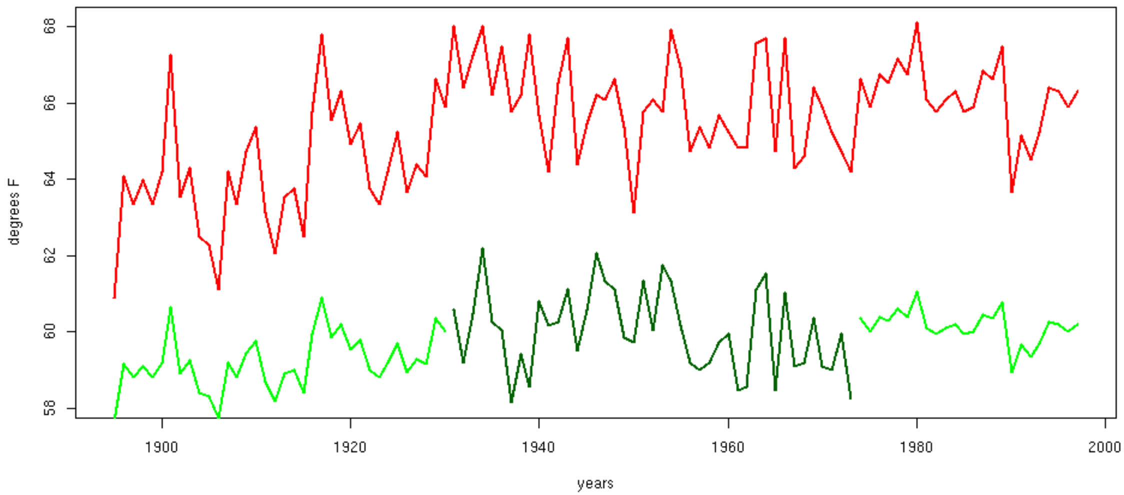
Connection with Least Squares (LS) If we use the sample statistics the conditional mean for Frasier is identical to

- Fitting a linear regression to the observed data.
- Using the LS line to predict a new temperature.

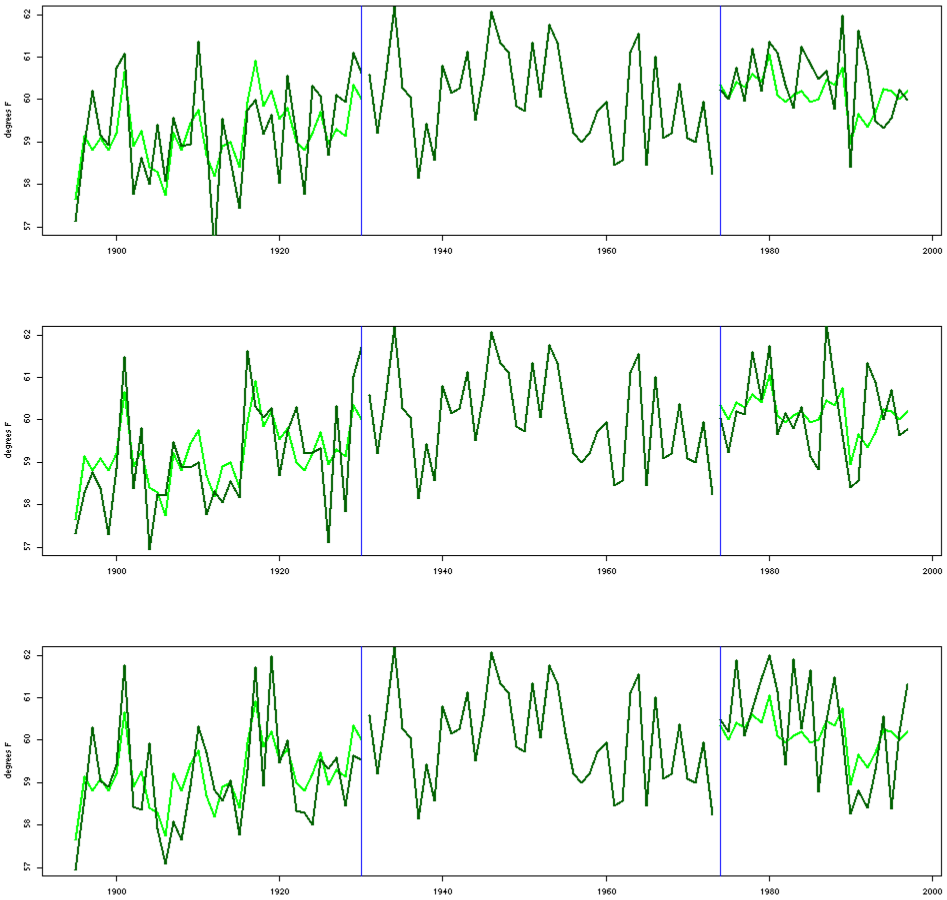
Connection with forecast skill The variance of the distribution gives a measure of the uncertainty in the prediction.

Analysis is only as good as the statistical assumptions!

Infilled Fraser means based on Boulder



Three members of an ensemble for Fraser



All infills have the same conditional mean and the variability will reproduce the climatology.

Spatial Statistics

The notorious "data product" What does the temperature field look like on a grid based on the observed data?

The model
$\mathbf{T}$ are the field values (e.g. temperatures) on a large, regular 2-d grid (and stacked as a vector). This is our universe.

$\mathbf{T}$ is multivariate normal with mean μ and covariance matrix: $\Sigma = COV(\mathbf{T})$ usually Σ is related to the distance between locations

The data
$\mathbf{Y}$ is the data taken at irregular locations

$$\mathbf{Y} = H\mathbf{T} + \mathbf{e}$$

$\mathbf{e}$ is measurement error, H is a known matrix that relates what we measure, on the average, to the true temperature field. In our case H is just an indicator matrix of ones and zeroes.

Kriging solution

$$\hat{\mathbf{T}} = \mu + COV(\mathbf{T}, \mathbf{Y})COV(\mathbf{Y})^{-1}(\mathbf{Y} - H\mu)$$

and the covariance of the estimate is

$$P = COV(\mathbf{T}) - COV(\mathbf{T}, \mathbf{Y})COV(\mathbf{Y})COV(\mathbf{Y}, \mathbf{T})$$

Bayesian solution

likelihood: data "given" temperature field = $[Y|T]$

prior: distribution of temperature field = $[T]$

Using *Bayes Theorem*

posterior: the conditional distribution of the temperatures "given" the data

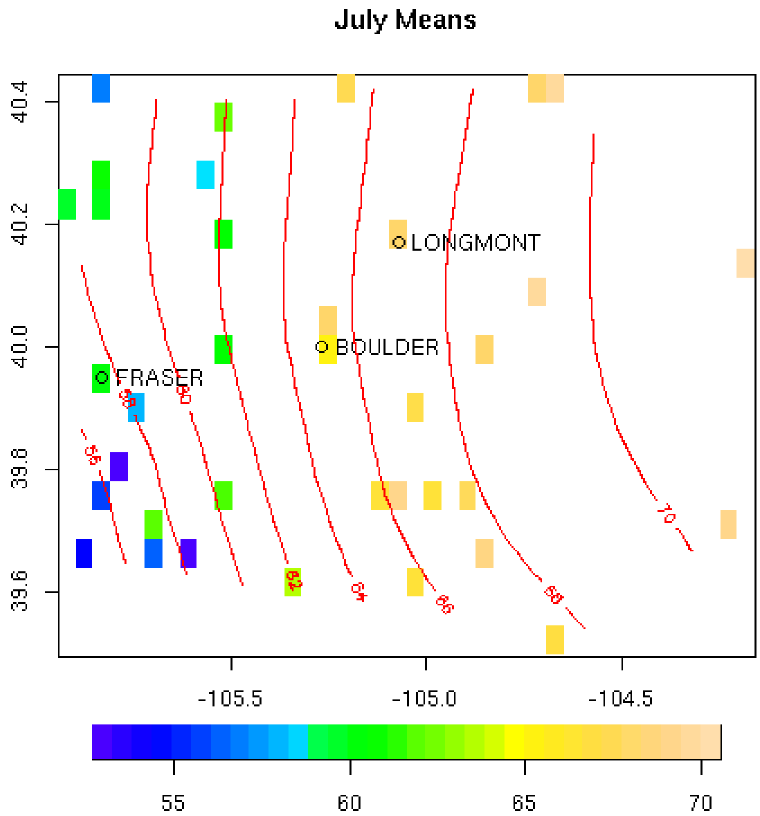
$$[T|Y] = \frac{[Y|T][T]}{[Y]}$$

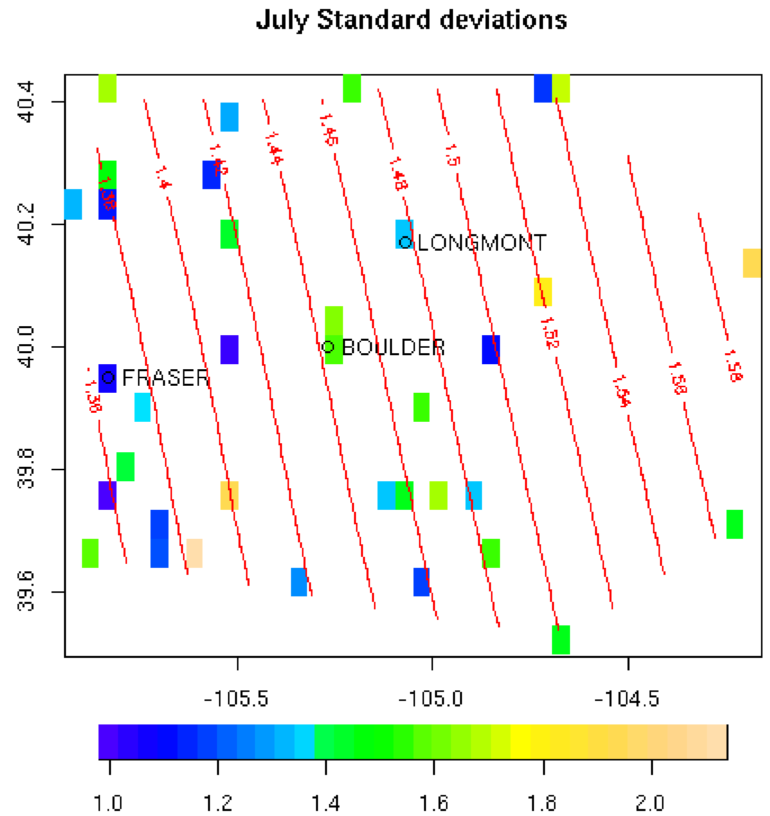
Posterior temperature field given the data is multivariate normal with mean vector $\hat{\mathbf{T}}$ and covariance matrix P!

Temperature fields for the Front Range

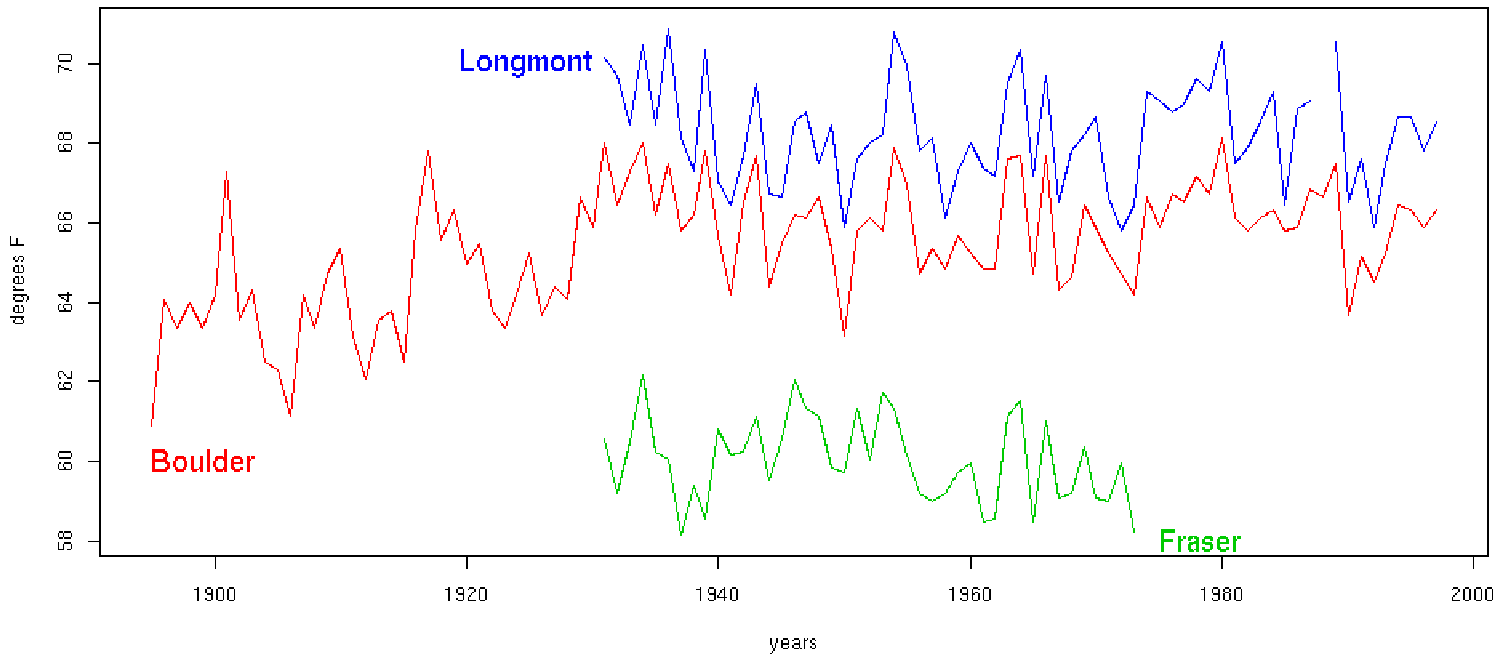
Estimating the means, variances and and correlations

μ and Σ for $\mathbf{T}$ are estimated from what data we have.

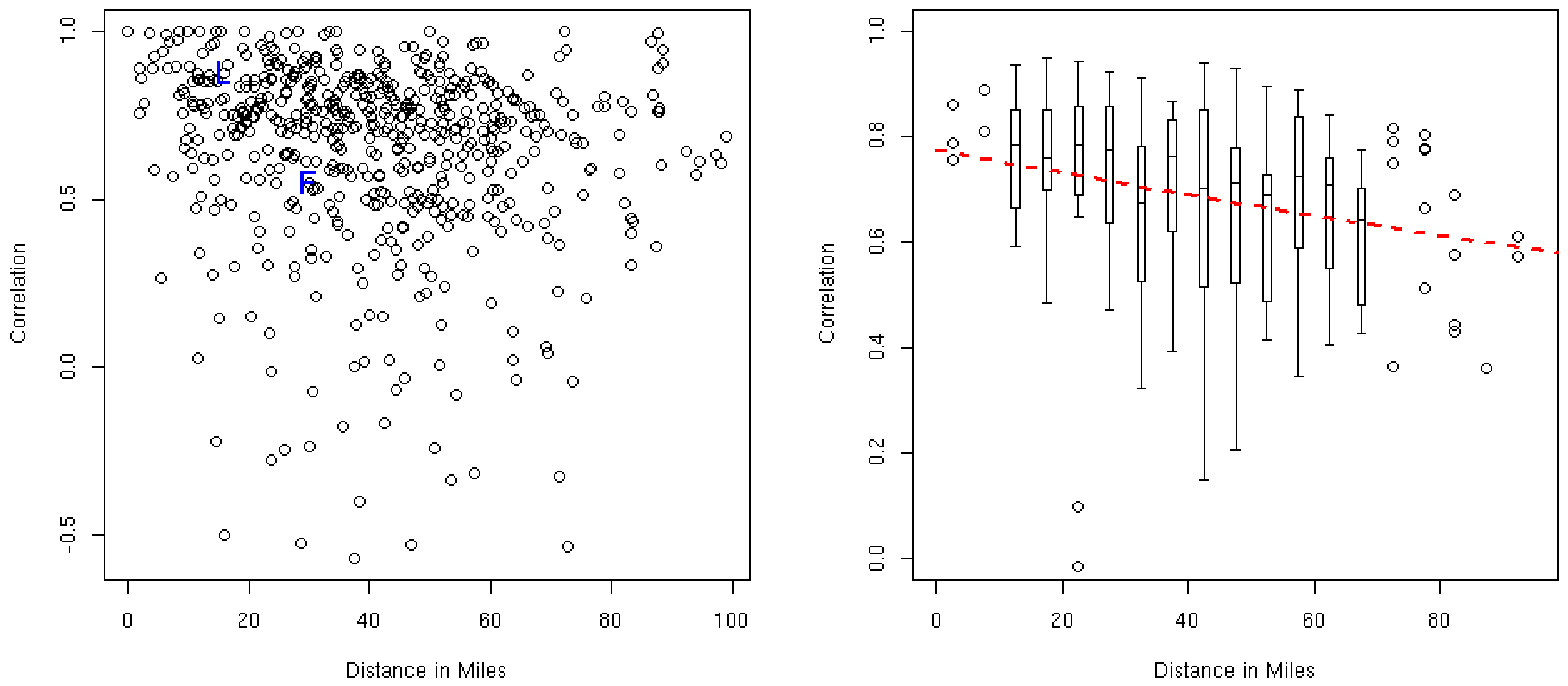




Spatial correlation of temperature

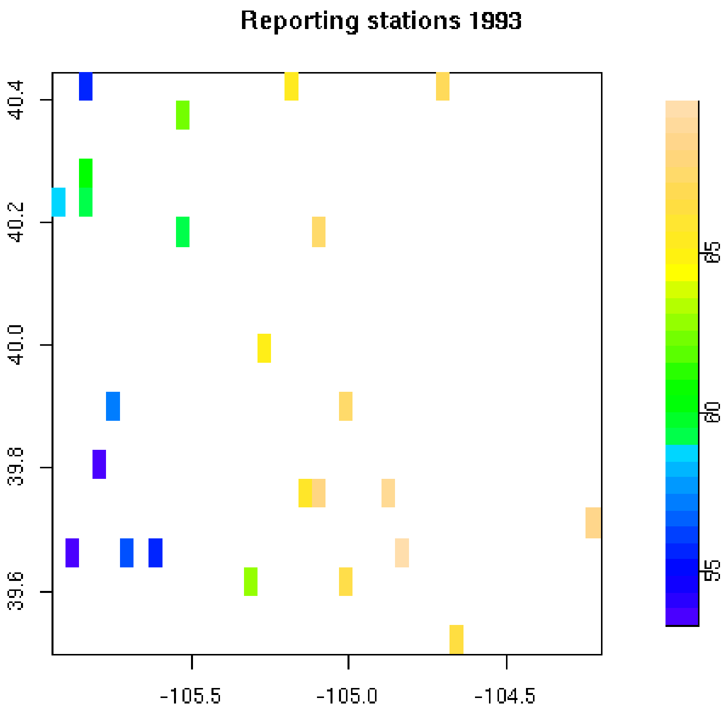


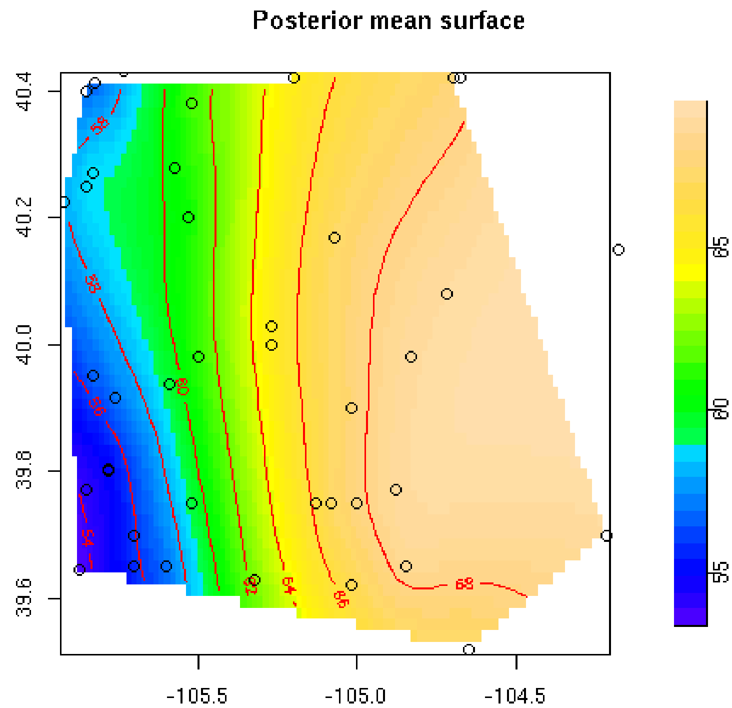
Dependence of correlation on distance



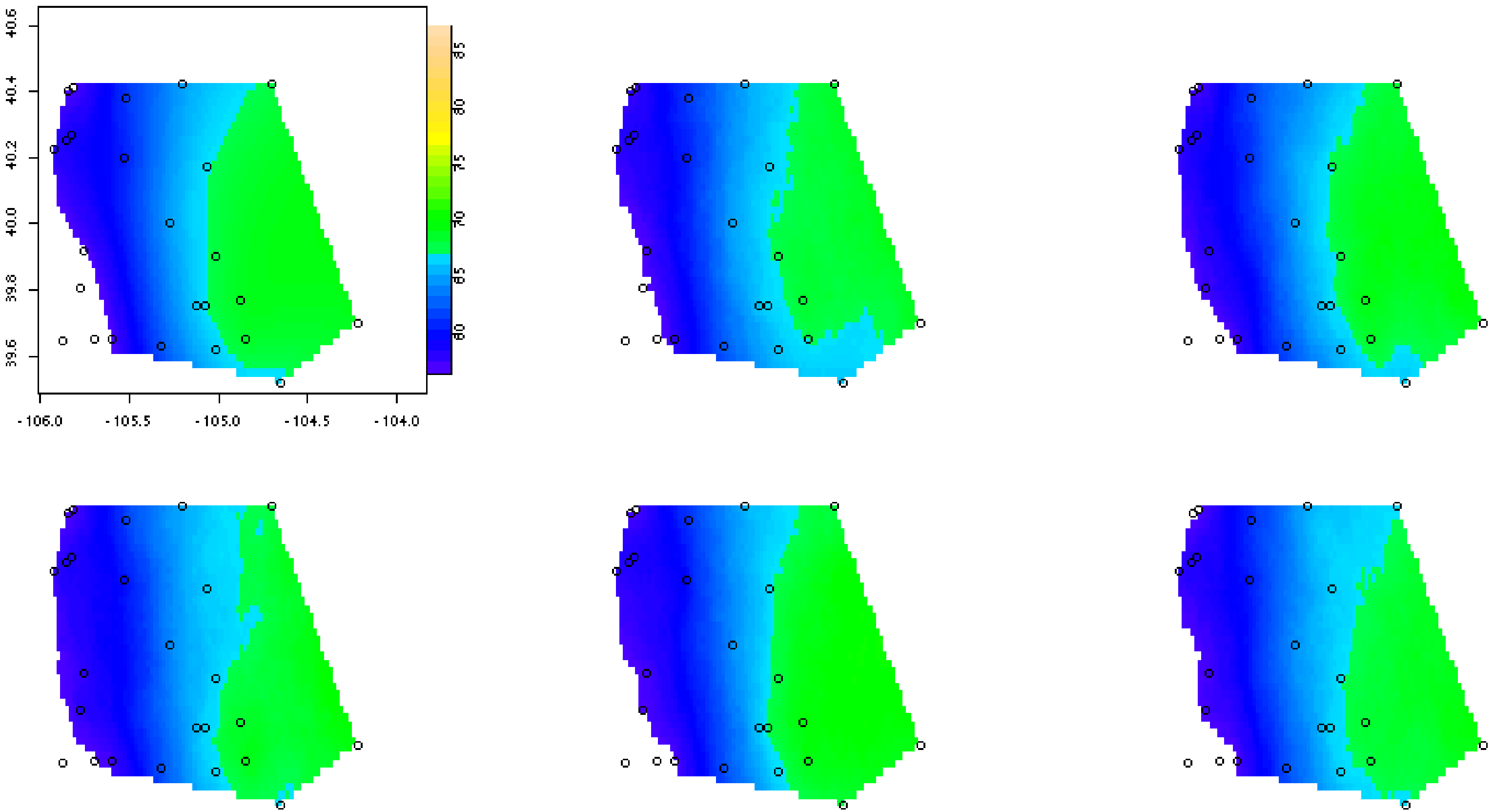
Note that the correlation is not zero close to zero distance! This may be due to measurement error.

Example of a posterior mean





Ensemble of fields for July 1993



Summary

- pdf can be approximated by samples
- conditional distributions can be predictive
- spatial prediction with observation error is an application of Bayes theorem.