# Conditional Expectation and Prediction

Statistics 110

Summer 2006

# Conditional Expectation

**Definition.** *The* **Conditional Expectation** *of* $Y$ *given* $X = x$ *is*

$$E[Y|X = x] = \begin{cases} \sum_y yp(y|x) & \textit{Discrete RV} \\ \int_{\mathcal{Y}} yf(y|x)dy & \textit{Continuous RV} \end{cases}$$

*More generally (for the continuous example),*

$$E[h(Y)|X = x] = \int_{\mathcal{Y}} h(y)f(y|x)dy$$

The conditional variance is given by

$$\begin{aligned} \mathrm{Var}(Y|X = x) &= E[(Y - E[Y|X = x])^2|X = x] \\ &= E[Y^2|X = x] - (E[Y|X = x])^2 \end{aligned}$$

Notice that all we are doing with conditional expectations is the standard calculations with the conditional distribution.

Example:

$$f(x,y) = \frac{1}{y^2} e^{-x/y^2} e^{-y}; \qquad x \geq 0, y > 0$$

so

$$f(y) = e^{-y}$$

$$f(x|y) = \frac{1}{y^2} e^{-x/y^2} \qquad (X|Y = y \sim Exp(1/y^2))$$

Therefore

$$E[X|Y = y] = \frac{1}{1/y^2} = y^2$$

$$\mathrm{Var}(X|Y = y) = \frac{1}{(1/y^2)^2} = y^4$$

Note that for any $h$, $E[h(Y)|X = x]$ is a function of $x$ (say $H(x)$). Since $X$ is a random variable, so is $H(X)$. So we can talk about their expectation and variance.

Of particular interest are

$$g(X) = E[Y|X]$$

and

$$h(X) = \mathrm{Var}(Y|X)$$

There are two important theorems about these quantities

**Theorem.** *Iterated Expectation*

$$E[E[X|Y]] = E[X]$$

**Proof.** Let $g(y) = E[X|Y = y]$

$$
\begin{aligned}
E[g(Y)] &= \int g(y) f_Y(y) dy \qquad \text{(Assume continuous)} \\
&= \int \left( \int x f_{X|Y}(x|y) dx \right) f_Y(y) dy \\
&= \int \int x \frac{f_{X,Y}(x,y)}{f_Y(y)} f_Y(y) dx dy \\
&= \int \int x f_{X,Y}(x,y) dy dx = E[X]
\end{aligned}
$$

□

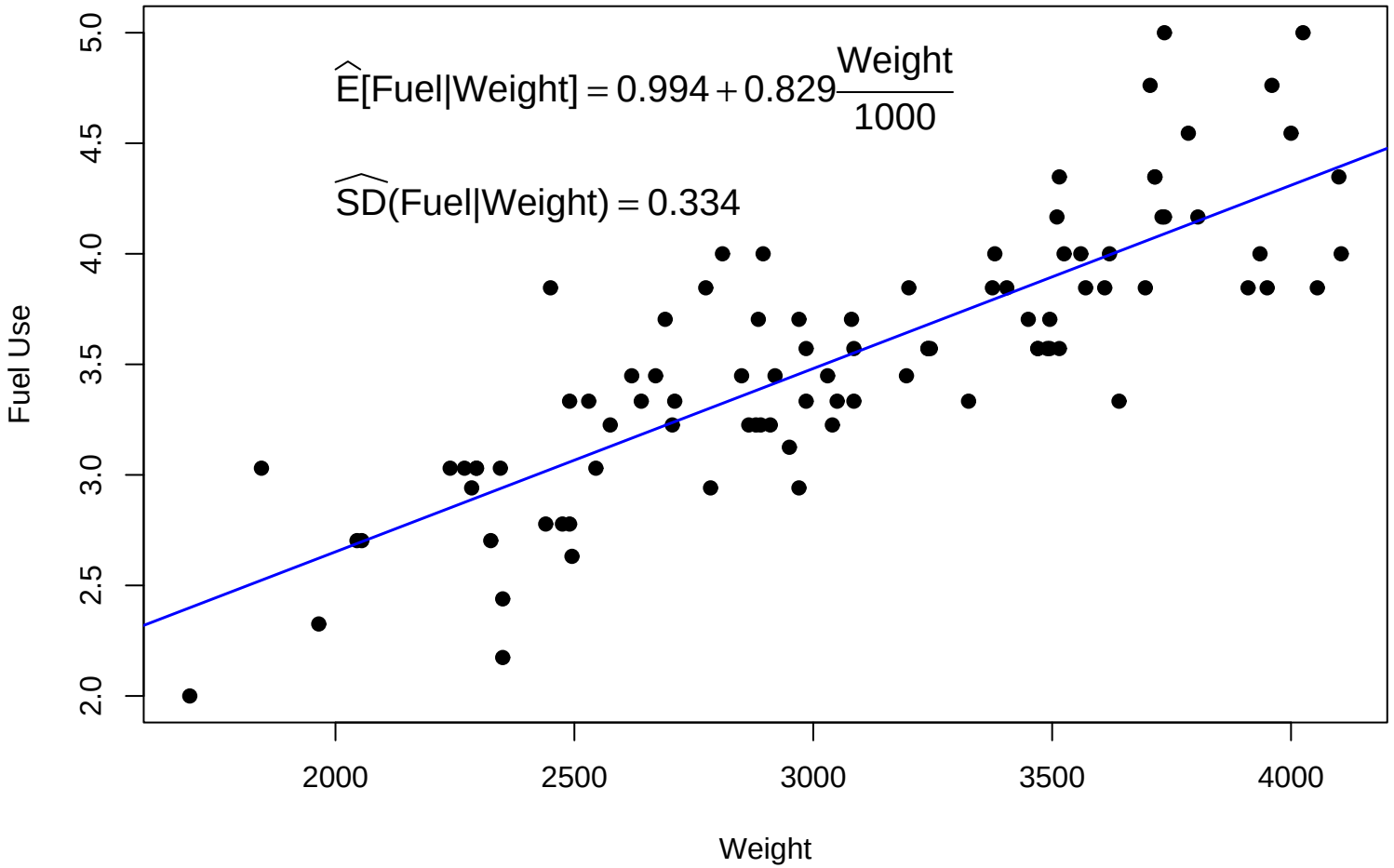
For the example, $E[X|Y] = Y^2$, $f_Y(y) = e^{-y}$

$$\begin{aligned} E[X] &= E[E[X|Y]] = E[Y^2] \\ &= \int_0^\infty y^2 e^{-y} dy = \Gamma(3) = 2! = 2 \end{aligned}$$

This theorem can be thought of as a law of total expectation. The expectation of a RV $X$ can be calculated by weighting the conditional expectations appropriately and summing or integrating.

## Example: Fuel Use

$X =$ Car Weight, $\quad Y = \frac{100}{\text{MPG}}$ (Gallons to go 100 miles)



Model for Fuel Use: $Y|X = x \sim N(\alpha + \beta x, \sigma^2)$

Suppose we want to get a handle the marginal distribution of fuel use. This depends on the breakdown of the weight of cars.

If there are more heavy cars, the overall fuel use should be higher.

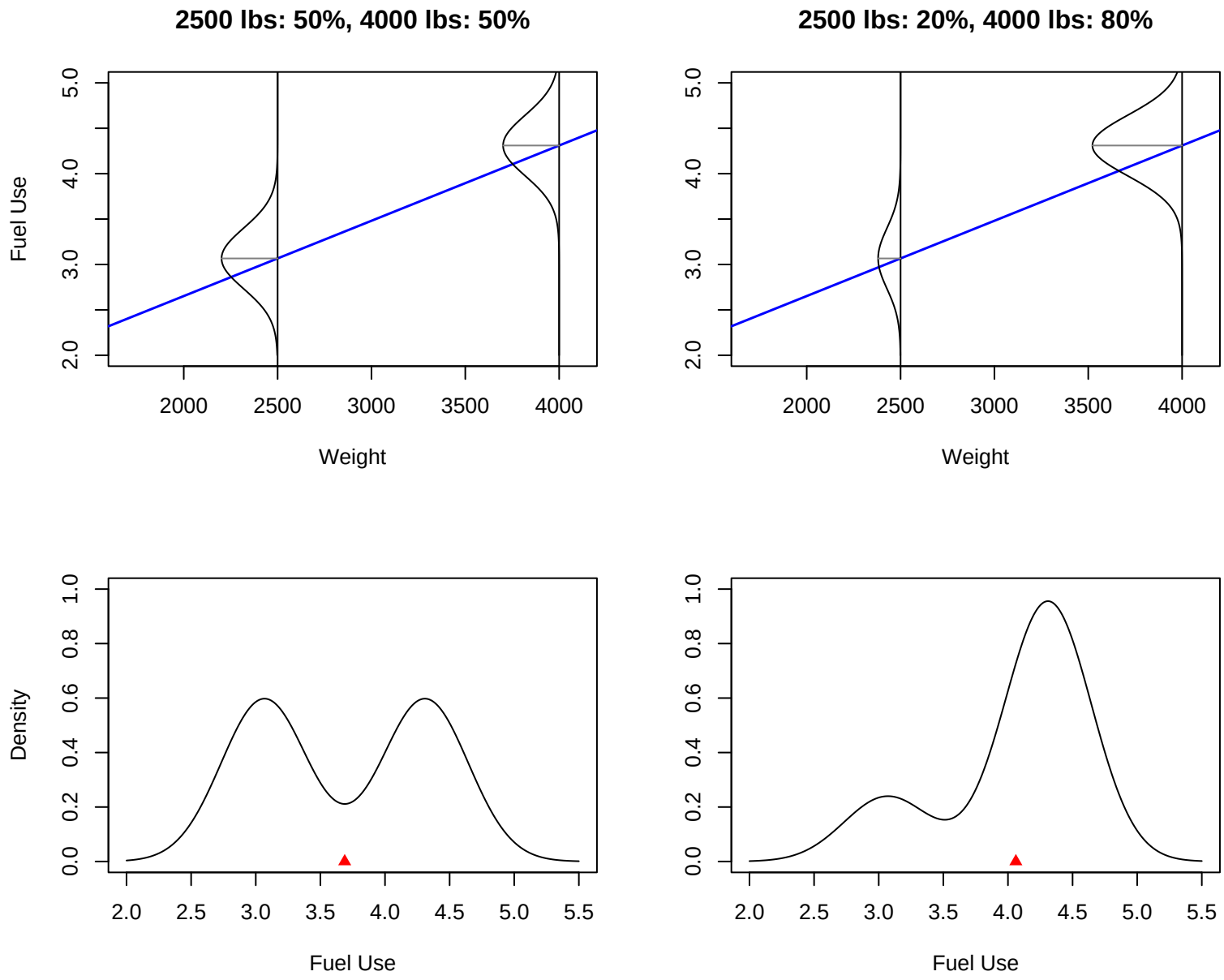
Lets consider two situations, both dealing with only 2500 lbs cars (mean = 3.067 gal) and 4000 lbs cars (mean = 4.310 gal).

1. 2500 lbs: 50%, 4000 lbs: 50%

$$E[\text{Fuel}] = 0.5 \times 3.067 + 0.5 \times 4.310 = 3.688$$

2. 2500 lbs: 20%, 4000 lbs: 80%

$$E[\text{Fuel}] = 0.2 \times 3.067 + 0.8 \times 4.310 = 4.061$$


2500 lbs: 50%, 4000 lbs: 50%
2500 lbs: 20%, 4000 lbs: 80%
Fuel Use
Weight
Density
2.0
3.0
4.0
5.0
2000
2500
3000
3500
4000
0.0
0.2
0.4
0.6
0.8
1.0
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5

In the survey response example discussed earlier

$$N \sim Bin(M, \pi)$$
$$X|N = n \sim Bin(n, p)$$

So $E[X]$, the expected number of people participating in the survey satisfies

$$E[X] = E[E[X|N]] = E[Np] = pE[N] = pM\pi$$

or by doing the algebra

$$= \sum_{n=0}^{M} np \binom{M}{n} \pi^n (1-\pi)^{M-n}$$
$$= p \sum_{n=0}^{M} n \binom{M}{n} \pi^n (1-\pi)^{M-n} = pM\pi$$

**Theorem.** *Variance Decomposition*

$$\mathrm{Var}(X) = \mathrm{Var}(E[X|Y]) + E[\mathrm{Var}(X|Y)]$$

*i.e.*

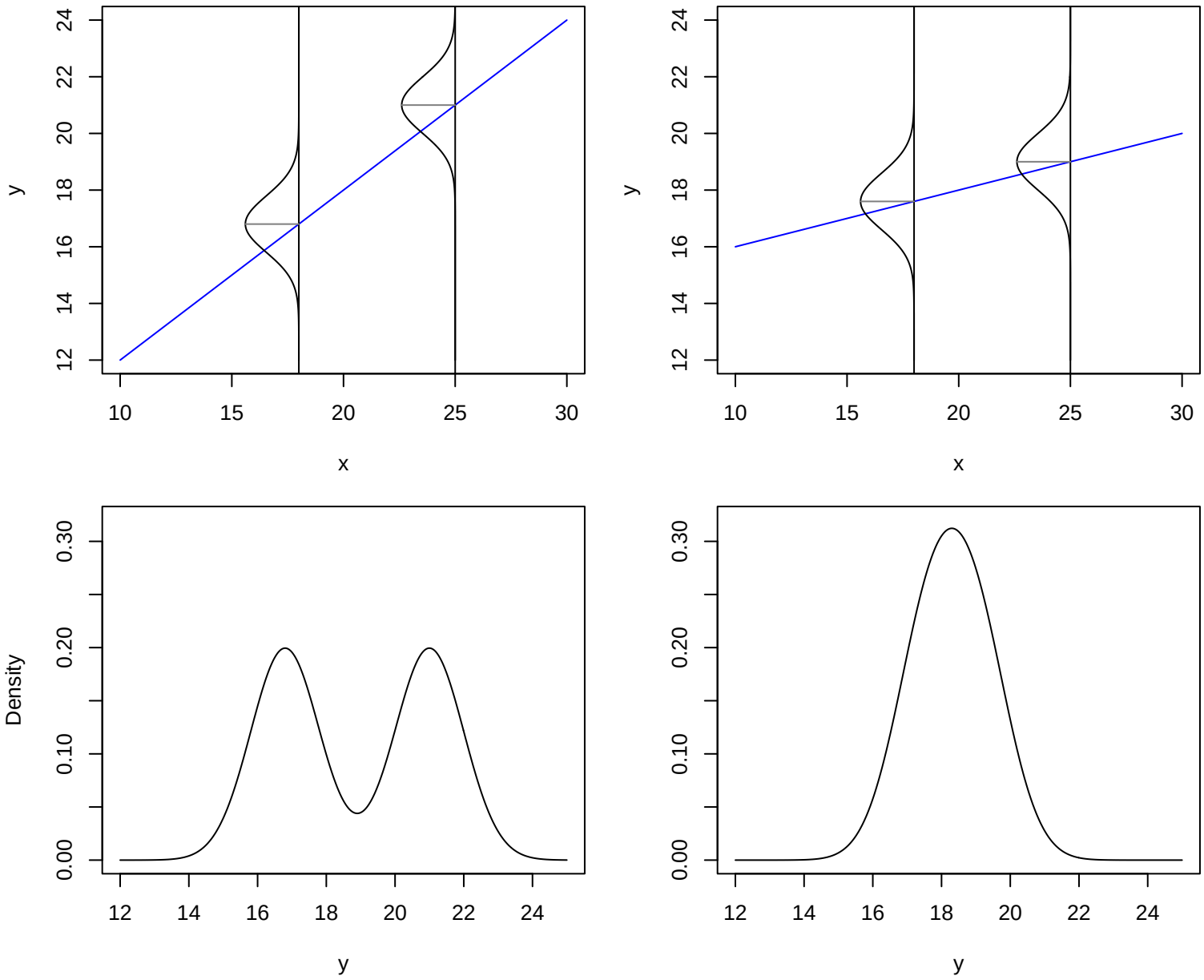
$$\mathrm{Var}(X) = \mathrm{Var}(g(Y)) + E[h(Y)]$$

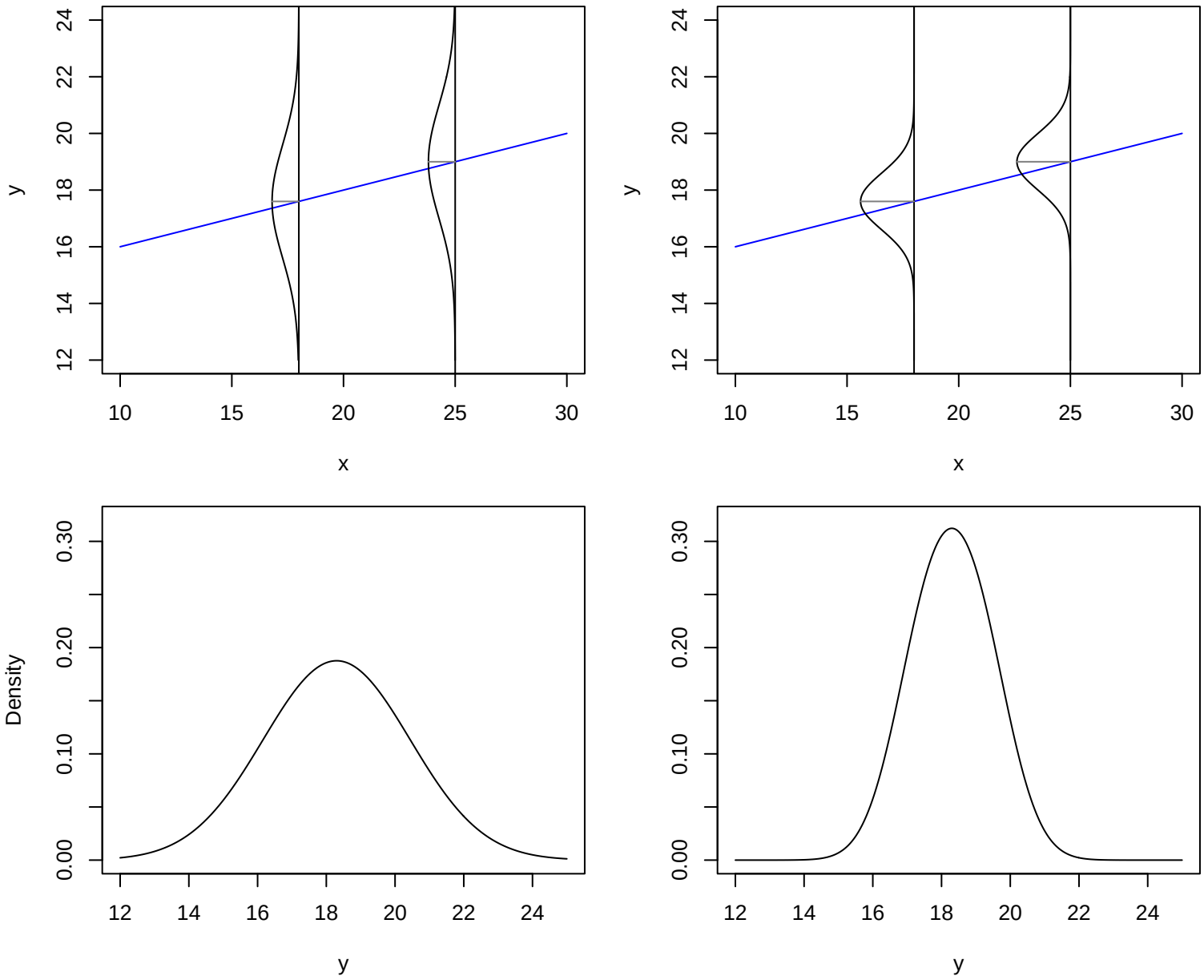
What this result is implying, when considering the spread of a random variable in the presence of another random variable (say a grouping variable), their are two important factors

1. How spread out are the means of the different groups – $\mathrm{Var}(E[X|Y])$ term

2. How spread out are the observations within each group – $E[\mathrm{Var}(X|Y)]$ term

(This decomposition underlies Analysis of Variance (ANOVA))


y
x
Density
y
x
y


y
x
Density
y
x
y

**Proof.**

$$\mathrm{Var}(X|Y=y) = E[X^2|Y=y] - (E[X|Y=y])^2$$

so

$$h(y) = E[X^2|Y=y] - (g(y))^2$$

$$\begin{aligned} E[h(Y)] = & \ E[E[X^2|Y]] - E[(g(Y))^2] \\ &= E[X^2] - \left(\mathrm{Var}(g(Y)) + (E[g(Y)])^2\right) \\ &= E[X^2] - \mathrm{Var}(g(Y)) - (E[X])^2 \\ &= \mathrm{Var}(X) - \mathrm{Var}(g(Y)) \end{aligned}$$

□

Back to exponential example ($E[X|Y] = Y^2, \mathrm{Var}(X|Y) = Y^4$)

$$\begin{aligned}
\mathrm{Var}(X) &= E[\mathrm{Var}(X|Y)] + \mathrm{Var}(E[X|Y]) \\
&= E[Y^4] + \mathrm{Var}(Y^2) \\
&= E[Y^4] + (E[Y^4] - (E[Y^2])^2) \\
&= 2 \times 4! - 2^2 = 44
\end{aligned}$$

since

$$E[Y^k] = \int_0^\infty y^k e^{-y} dy = \Gamma(k+1) = k!$$

Binomial Example ($E[X|N] = Np, \mathrm{Var}(X|N) = Np(1-p), E[N] = M\pi, \mathrm{Var}(N) = M\pi(1-\pi)$)

$$\begin{aligned}
\mathrm{Var}(X) &= E[\mathrm{Var}(X|N)] + \mathrm{Var}(E[X|N]) \\
&= E[Np(1-p)] + \mathrm{Var}(Np) \\
&= p(1-p)E[N] + p^2\mathrm{Var}(N) \\
&= p(1-p)M\pi + p^2M\pi(1-\pi) \\
&= p\pi M - p^2\pi M + p^2\pi M - p^2\pi^2 M = Mp\pi(1-p\pi)
\end{aligned}$$

Actually we already knew this result since we've shown that $X \sim Bin(M, p\pi)$

These two results can make difficult moment calculations easy to do. For example, the initial example

$$f(x,y) = \frac{1}{y^2} e^{-x/y^2} e^{-y}; \quad x \geq 0, y > 0$$

so

$$f(y) = e^{-y}$$
$$f(x|y) = \frac{1}{y^2} e^{-x/y^2} \qquad (X|Y = y \sim Exp(1/y^2))$$

getting the marginal density of $X$ is not easy (its absolutely ugly).

Even though we couldn't calculate the integrals directly, we can still determine the moments of the marginal distribution.

They also allow us to think in terms of hierarchical models, building pieces one on top of the other.

Note that the examples so far have either been all discrete RVs or all continuous RVs. There is no reason to restrict to these cases. You can have a mixture of continuous and discrete RVs.

For example, a more specific case of the random sums (example D on page 138) would be

$$\begin{aligned} N &\sim Pois(\mu) \\ T|N = n &\sim \sum_{i=1}^{n} X_i \qquad \text{where } X_i \sim Gamma(\alpha, \lambda) \\ &\sim Gamma(n\alpha, \lambda) \end{aligned}$$

So

$$E[T] = E[E[T|N]] = E\left[N\frac{\alpha}{\lambda}\right] = \mu\frac{\alpha}{\lambda}$$

$$\begin{aligned}
\mathrm{Var}(T) &= \mathrm{Var}(E[T|N]) + E[\mathrm{Var}(T|N)] \\
&= \mathrm{Var}\left(N\frac{\alpha}{\lambda}\right) + E\left[N\frac{\alpha}{\lambda^2}\right] \\
&= \frac{\alpha^2}{\lambda^2}\mathrm{Var}(N) + \frac{\alpha}{\lambda^2}E[N] \\
&= \frac{\alpha^2}{\lambda^2}\mu + \frac{\alpha}{\lambda^2}\mu = \mu\frac{\alpha}{\lambda^2}(\alpha+1)
\end{aligned}$$

The factor $\alpha + 1$ tells us how much the variance gets increased due to our lack of knowledge of $N$, the number of terms summed.

In this example, the conditioning variable was discrete and the variable of interest was continuous. Note that we can go the other way as well.

$$\begin{aligned}
\lambda &\sim Exp(\mu) \\
X|\lambda &\sim Pois(\lambda)
\end{aligned}$$

This model comes about in the situations that we expect that a count should have a Poisson distribution, but we aren't sure of the rate. So we can describe our uncertainty about the rate with a probability distribution. One choice is a exponential distribution (Gamma is a more popular choice).

$$E[\lambda] = \frac{1}{\mu}; \mathrm{Var}(\lambda) = \frac{1}{\mu^2}$$

$$E[X] = E[E[X|\lambda]] = E[\lambda] = \frac{1}{\mu}$$

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(E[X|\lambda]) + E[\mathrm{Var}(X|\lambda)] \\ &= \mathrm{Var}(\lambda) + E[\lambda] \\ &= \frac{1}{\mu^2} + \frac{1}{\mu}\end{aligned}$$

The extra $\frac{1}{\lambda^2}$ term is the extra uncertainty in $X$ due to not knowing the exact mean of the Poisson distribution.

Note that in these situations, we can figure out the marginal and conditional distribution that aren't given. For the second Poisson/Gamma example, the joint "density" is given by

$$f_{X,\lambda}(x,\lambda) = p_{X|\lambda}(x|\lambda) f_\lambda(\lambda); \quad x = 0, 1, 2, \ldots, \lambda > 0$$

So the marginal PMF of $X$ is given by

$$\begin{aligned} p_X(x) &= \int_0^\infty f_{X,\lambda}(x,\lambda) d\lambda \\ &= \int_0^\infty \mu \frac{\lambda^x}{\Gamma(x+1)} e^{-\lambda(1+\mu)} d\lambda \\ &= \frac{\mu}{(1+\mu)^{x+1}}; \quad x = 0, 1, 2, \ldots \end{aligned}$$

(Aside: Note that this distribution is related to the Geometric distribution with success probability $\frac{\mu}{1+\mu}$. Here $x$ would correspond to the number of "failures" before the first "success".)

and the conditional density of $\lambda|X = x$ is

$$f_{\lambda|X}(\lambda|x) = \frac{f_{X,\lambda}(x,\lambda)}{p_X(x)} = \frac{\lambda^x(1+\mu)^{x+1}}{\Gamma(x+1)}e^{-\lambda(1+\mu)}; \quad \lambda > 0$$

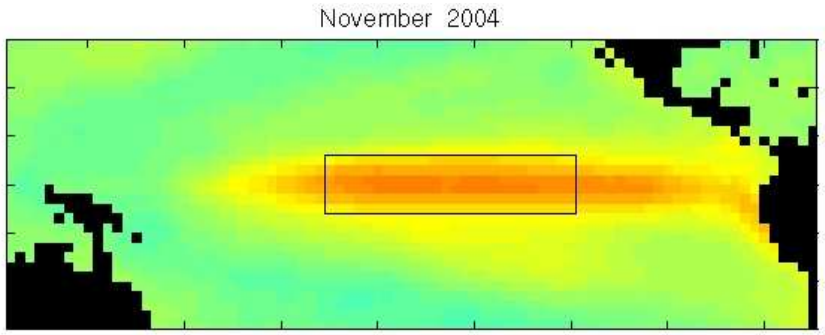
so $\lambda|X = x \sim Gamma(x+1, \mu+1)$

# Optimal Prediction

A probability distribution gives a measure of knowledge or believe about a random process of interest. However in many situations it is often useful to be able to come up with a single prediction of what we might observe if we were to generate a new realization of the process.

Examples:

- In the SST example, the model gives us a probability distribution for the temperature at different locations in the tropical Pacific. For forecasting purposes it is useful to have a single temperature prediction for each location



- Uncertain binomial success probabilities

  We want to sample from a population consisting of two type of members (John McCain voters and Hilary Clinton voters). However the fraction of the two types is unknown ($p$: fraction of McCain voters, $q = 1 - p$: fraction of Clinton voters). So we can take a sample of size $n$ from the population to learn about $p$ and $q$. Suppose that we have a prior belief about what $p$ might be given in the form of a probability distribution.

$$X|p \sim Bin(n, p)$$
$$p \sim Beta(a, b) \qquad \text{(Prior belief)}$$

  We want to use the observed data $x$ and the prior belief to come up with our best guess for $p$.

<aside> The joint "density" of $X$ and $p$ is

$$f_{X,p}(x,p) = \binom{n}{x} p^x (1-p)^{n-x} \times \frac{1}{\beta(a,b)} p^{a-1}(1-p)^{b-1};$$
$$x = 0, 1, \ldots, n, 0 < p < 1$$

The marginal PMF of $X$ is

$$p_X(x) = \binom{n}{x} \frac{\beta(a+x, b+n-x)}{\beta(a,b)}$$

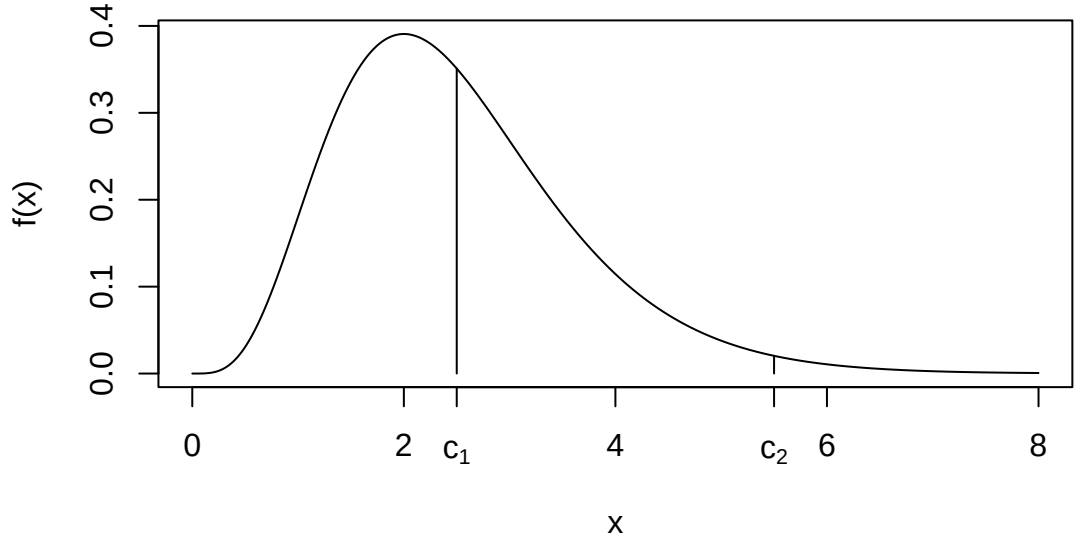
This is known as the Beta-Binomial distribution. The conditional density of $p|X = x$ is

$$f_{p|X}(p|x) = \frac{1}{\beta(a+x, b+n-x)} p^{a+x-1}(1-p)^{b+n-x-1}; \quad 0 < p < 1$$

i.e, $Y|X = x \sim Beta(a + x, b + n - x)$. </aside>

What should be use as a predictor? We want something that picks something that is close to values of the random variable $Y$ that are highly probable. We need a criterion that measures how well we do if our prediction is the value $c$. A popular choice is the mean squared error (MSE)

$$MSE(c) = E[(Y - c)^2]$$



**Theorem.** *Under the MSE criterion, the optimal predictor is* $c = E[Y]$.

**Proof.**

$$E[(Y-c)^2] = \mathrm{Var}(Y-c) + (E[Y-c])^2 = \mathrm{Var}(Y) + (\mu - c)^2$$

The variance piece doesn't depend on our choice of predictor, so we only need to minimize the second term, which is done by setting $c = \mu = E[Y]$. $\square$

Note that we often have other information available that we want to include.

- In the SST example, we have the past temperatures, the wind and pressure data.
- In the sampling example, we have the poll data $X$.

So in this case, we need to choose a function $h(X)$ to minimize $MSE(h(X)) = E[(Y - h(X))^2]$

Note that

$$E[(Y - h(X))^2] = E[E[(Y - h(X))^2|X]]$$

So by the previous theorem, for each fixed $x$, the inner expectation $E[(Y - h(X))^2|X = x]$ is minimized by $h(x) = E[Y|X = x]$, thus the minimizing function is

$$h(X) = E[Y|X]$$

So for the polling example, our best guess for $p$ is $\frac{a+x}{a+b+n}$, which is the mean of a $Beta(a + x, b + n - x)$. Note that this happens to be between the forecast based on the prior $\frac{a}{a+b}$ and the sample proportion of McCain supporters $\frac{x}{n}$.

Now let $X$ and $Y$ be bivariate normal. Then

$$E[Y|X] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X) = \alpha + \beta X$$

Now in many problems, the conditional mean of $Y|X = x$ can be difficult to work with. So instead of trying to find the best function, lets try to find a function from a restricted class, such as linear predictors ($h(x) = \alpha + \beta x$).

Want to find the best choices for $\alpha$ and $\beta$, that is choose them to minimize

$$MSE(\alpha, \beta) = E[(Y - (\alpha + \beta X))^2]$$

One way to choose them would be the find the gradient, set it to $0$, and solve. Instead we can do it another way that doesn't need multivariate calculus.

$$\begin{aligned} E[(Y - (\alpha + \beta X))^2] &= \mathrm{Var}(Y - (\alpha + \beta X)) + (E[Y - (\alpha + \beta X)])^2 \\ &= \mathrm{Var}(Y - \beta X) + (E[Y - (\alpha + \beta X)])^2 \end{aligned}$$

As the first term doesn't depend on $\alpha$, we can figure out what the best choice for it is for each possible $\beta$, and then get the best $\beta$.

Note that the second term can be made to be zero by setting

$$\alpha = \mu_Y - \beta\mu_X$$

The first term is

$$\mathrm{Var}(Y - \beta X) = \sigma_Y^2 + \beta^2\sigma_X^2 - 2\beta\sigma_{XY}$$

where $\sigma_{XY} = \mathrm{Cov}(X, Y)$. This variance is minimized by setting

$$\beta = \frac{\sigma_{XY}}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X}$$

Plugging these values in $\alpha$ and $\beta$ gives the minimum mean squared error linear predictor

$$\hat{Y} = \alpha + \beta X = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)$$

Note that for this linear predictor, we don't need to know the complete conditional distribution. Instead we need to know the marginal means and variances, and the correlation (or covariance).

Note that this result supports the idea that the correlation is a measure of the strength of the linear relationship between two variables.

While we are looking for a single prediction of the random variable of interest, it is useful to also have a measure of uncertainty about that prediction. The usual choice is the variance, as for the optimal predictor

$$MSE(h(X)) = \mathrm{Var}(Y - h(X)) = E[\mathrm{Var}(Y|X)]$$

For the linear predictor

$$\begin{aligned}\mathrm{Var}(Y - \beta X) &= \sigma_Y^2 + \frac{\sigma_{XY}^2}{\sigma_X^4}\sigma_X^2 - 2\frac{\sigma_{XY}}{\sigma_X^2}\sigma_{XY} \\ &= \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2} \\ &= \sigma_Y^2 - \rho^2\sigma_Y^2 = \sigma_Y^2(1-\rho^2)\end{aligned}$$

Again, this doesn't depend on the conditional distribution, but only the first two moments of $X$ and $Y$. Note that this is $\mathrm{Var}(Y|X)$ if $X$ and $Y$ are bivariate normal, which is to be expected as the linear predictor is also the optimal predictor in that case.

These mathematical arguments help support the wide use of linear regression techniques for many problems.

Note that other optimality criteria can be used. For example, the Mean Absolute Deviation (MAD)

$$MAD(h(X)) = E[|Y - h(X)|]$$

leads to $h(x)$ being the median of the distribution of $Y|X = x$.

What's known as 0–1 loss leads to the mode of the distribution of $Y|X = x$, the $y$ with the largest density (continuous) or probability (discrete).

These tend to be used less, as mathematically they are less tractable, particularly if you wish to restrict $h(x)$ to the class of linear predictors. For example, with the MAD criterion,

$$E[|Y - (\alpha + \beta X)|]$$

is difficult to optimize since the function $|x|$ is not differentiable at $0$.

Also the variance is not the best choice for our uncertainty measure of the predictor. Something based on a MAD type measure or an interquartile range would be more appropriate, though the variance is still of use.

Also the MSE based predictors have been shown to work well over time.

Also for some problems, linear predictors won't work well. For example suppose you want to predict a random variable restricted to the range (0,1). A linear predictor may not work well, as eventually $\alpha + \beta X$ must go outside the range (0,1). A possible predictor in that case could have the form

$$h(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$