Continuous Random Variables (Devore Chapter Four)

1016-351-01: Probability*

Fall 2011

Contents

Tuesday 11 October 2011

1 Continuous Random Variables

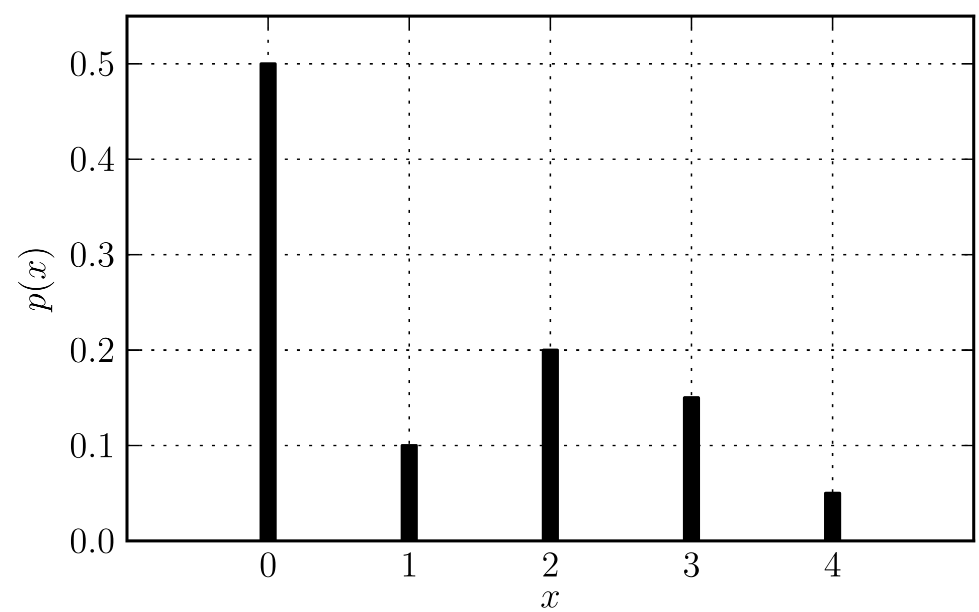
Recall that when we introduced the concept of random variables last chapter we focussed on discrete random variables, which could take on one of a set of specific values, either finite or countably infinite. Now we turn attention to **continuous random variables**, which can take on an value in one or more intervals, but for which there is zero probability for any single value.

1.1 Defining the Probability Density Function

Devore starts with the definition of the probability density function, and deduces the cumulative distribution function for a continuous random variable from that. We will follow a complementary presentation, starting by extending the cdf to a continuous rv, and then deriving the pdf from that.

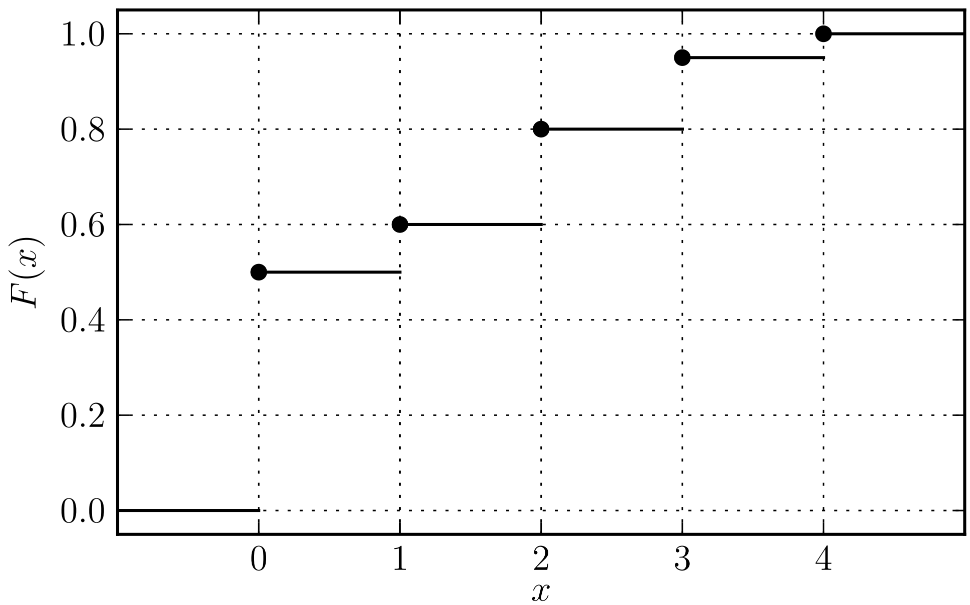
Recall that if X is a discrete random variable, we can define the probability mass function

$$p(x) = P(X = x) \tag{1.1}$$



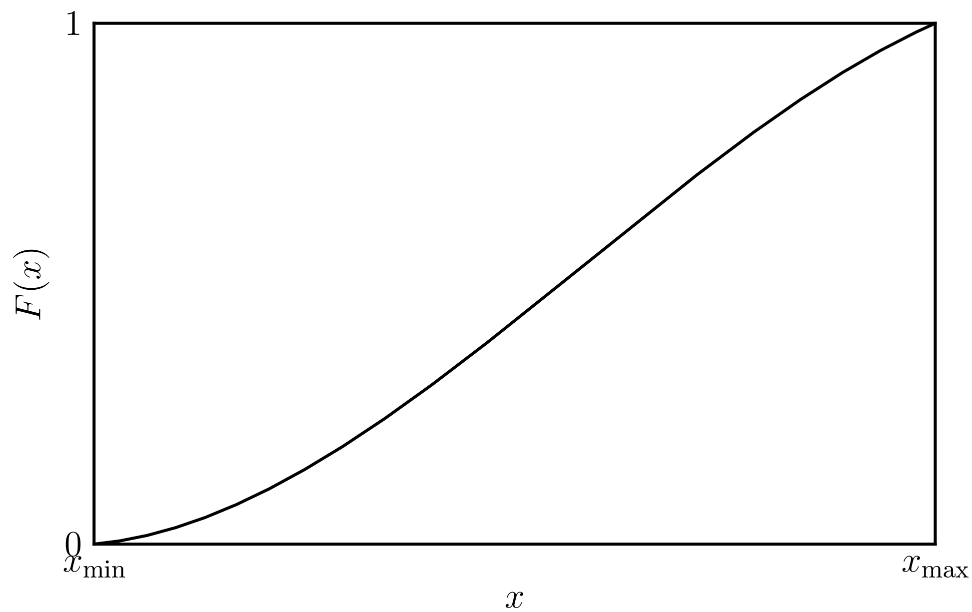
and the cumulative distribution function

$$F(x) = P(X \leq x) = \sum_{y \leq x} p(y) \tag{1.2}$$



Note that for a discrete random variable, the cdf is constant for a range of values, and then jumps when we reach one of the possible values for the rv.

For a continuous rv X, we know that the probability of X equalling exactly any specific x is zero, so clearly $p(x) = P(X = x)$ doesn't make any sense. However, $F(x) = P(X \leq x)$ is still a perfectly good definition. Now, though, $F(x)$ should be a continuous function with no "jumps", since a jump corresponds to a finite probability for some specific value:



Note that in both cases, if x_{min} and x_{max} are the minimum and maximum possible values of

X, then $F(x) = 0$ when $x < x_{\min}$, and $F(x) = 1$ when $x \geq x_{\max}$. We can also write

$$F(-\infty) = 0 \tag{1.3a}$$
$$F(\infty) = 1 \tag{1.3b}$$

Just as before, we can calculate the probability for X to lie in some finite interval from a to b, using the fact that

$$(X \leq b) = (X \leq a) \cup (a < X \leq b) \tag{1.4}$$

(where $b > a$):

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \ . \tag{1.5}$$

But now, since $P(X = a) = 0$ and $P(X = b) = 0$, we don't have to worry about whether or not the endpoints are included:

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) \qquad (X \text{ a cts rv}) \tag{1.6}$$

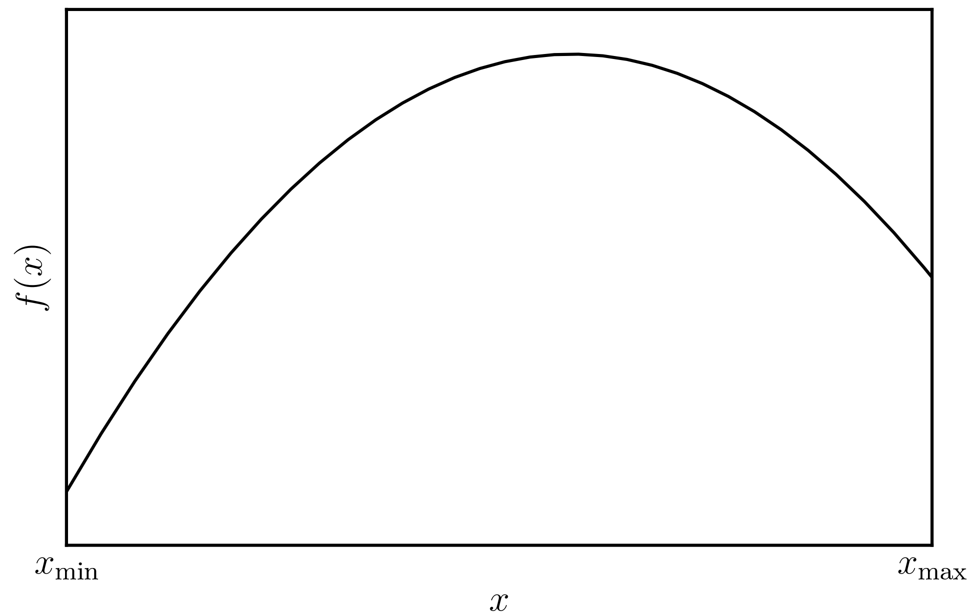
We can't talk about the probability for X to be exactly some value x, but we can consider the probability for X to lie in some little interval of width Δx centered at x:

$$P\left(x - \Delta x/2 < X < x + \Delta x/2\right) = F\left(x + \Delta x/2\right) - F\left(x - \Delta x/2\right) \tag{1.7}$$

We know that if we make Δx smaller and smaller, the difference between $F\left(x + \Delta x/2\right)$ and $F\left(x - \Delta x/2\right)$ has to get smaller and smaller. If we define

$$\begin{aligned} f(x) &= \lim_{\Delta x \to 0} \frac{P\left(x - \Delta x/2 < X < x + \Delta x/2\right)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{F\left(x + \Delta x/2\right) - F\left(x - \Delta x/2\right)}{\Delta x} = \frac{dF}{dx} = F'(x) \end{aligned} \tag{1.8}$$

this is just the derivative of the cdf. This derivative $f(x)$ is called the *probability density function*.



If we want to calculate the probability for X to lie in some interval, we can use the pdf $f(x)$ or the cdf $F(x)$. If the interval is small enough that the cdf $F(x)$ can be approximated as a straight line, the change in $F(x)$ over that interval is approximately the width of the interval times the derivative $f(x) = F'(x)$:

$$P\left(x - \Delta x/2 < X < x + \Delta x/2\right) = F\left(x + \Delta x/2\right) - F\left(x - \Delta x/2\right) \approx F'(x)\,\Delta x = f(x)\,\Delta x \tag{1.9}$$

If the interval is not small enough, we can always divide it up into pieces that are. If the ith piece is Δx wide and centered on x_i,

$$P\left(a < X < b\right) = F(b) - F(a) \approx \sum_{i=1}^{N} f(x_i)\,\Delta x \tag{1.10}$$

where

$$x_{i+i} - x_i = \Delta x \tag{1.11}$$

and

$$x_1 - \Delta x/2 = a \tag{1.12a}$$
$$x_N + \Delta x/2 = b \tag{1.12b}$$
$$\tag{1.12c}$$

In the limit that $N \to \infty$, so that $\Delta x \to 0$, the sum becomes an integral, so that

$$P\left(a < X < b\right) = F(b) - F(a) = \int_a^b f(x)\,dx \tag{1.13}$$

Of course, this is also a consequence of the Second Fundamental Theorem of Calculus:

$$\text{if } f(x) = F'(x) \text{ then } \int_a^b f(x)\,dx = F(b) - F(a) \tag{1.14}$$

The normalization condition, which for a discrete random variable is $\sum_x p(x) = 1$, becomes for a continuous random variable

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{1.15}$$

1.2 Expected Values

Recall that for a discrete random variable, the expected value is

$$E(X) = \sum_x x\,p(x) \qquad (X \text{ a discrete rv}) \tag{1.16}$$

In the case of a continuous random variable, we can use the same construction as before to divide the whole range from $x_{\min}$ to $x_{\max}$ into little intervals of width Δx, centered at points x_i, we can define a discrete random variable $X_{\Delta x}$ which is just X rounded off to middle of whatever interval it's in. The pmf for this discrete random variable is

$$p_{\Delta x}(x_i) = P\left(x_i - \Delta x/2 < X < x_i + \Delta x/2\right) \approx f(x_i)\,\Delta x \tag{1.17}$$

$X_{\Delta x}$ and X will differ by at most $\Delta x/2$, so we can write, approximately,

$$E(X) \approx E(X_{\Delta x}) = \sum_i x_i p_{\Delta x}(x_i) \approx \sum_i x_i f(x_i)\,\Delta x \tag{1.18}$$

In the limit that we divide up into more and more intervals and $\Delta x \to 0$, this becomes

$$E(X) = \int_{x_{\min}}^{x_{\max}} x\,f(x)\,dx = \int_{-\infty}^{\infty} x\,f(x)\,dx \tag{1.19}$$

The same argument works for the expected value of any function of X:

$$E(h(X)) = \int_{-\infty}^{\infty} h(x)\,f(x)\,dx \tag{1.20}$$

In particular,

$$E(X^2) = \int_{-\infty}^{\infty} x^2\,f(x)\,dx \tag{1.21}$$

We can also define the variance, and as usual, the same shortcut formula applies:

$$\sigma_X^2 = E\big((X - \mu_X)^2\big) = E(X^2) - \big(E(X)\big)^2 \tag{1.22}$$

where

$$\mu_X = E(X) \tag{1.23}$$

1.3 Percentiles and Medians

Another concept from descriptive statistics that has an analogue for random variables is that of percentiles, and specifically the 50th percentile or median.

Recall that the median was defined as the "middle" value in a data set, with half of the values above and half below it. We can define a similar concept for random variables: the median $\widetilde{\mu}$ of a probability distribution is the value such that there is a 50% chance of the random variable lying above the median, and a 50% chance of lying below it. We didn't talk about it much for discrete random variables, since a careful definition would be something like this:

$$P(X \geq \widetilde{\mu}) \geq .5 \qquad \text{and} \qquad P(X \leq \widetilde{\mu}) \geq .5 \tag{1.24}$$

With continuous random variables, it's a little easier. Since the probability of X equalling exactly $\widetilde{\mu}$ is zero, $P(X \geq \widetilde{\mu}) = P(X > \widetilde{\mu}) = 1 - P(X \leq \widetilde{\mu})$ and therefore

$$F(\widetilde{\mu}) = P(X \leq \widetilde{\mu}) = .5 \qquad \text{(for continuous random variables)} \tag{1.25}$$

I.e., the median of a probability distribution is the value at which the cumulative distribution function equals one-half.

Percentiles are defined similarly; the $(100 \times p)$th percentile, called $\eta(p)$, is the value at which the cdf equals p:

$$F(\eta(p)) = p \tag{1.26}$$

If we have the functional form of the cdf, we can just solve (1.26) for $\eta(p)$; if not, we can still define the percentile indirectly using the definition of the cdf:

$$\int_{-\infty}^{\eta(p)} f(x)\, dx = p \tag{1.27}$$

1.4 Worksheet

- worksheet
- solutions

Practice Problems

4.1, 4.5, 4.11, 4.13, 4.17, 4.21

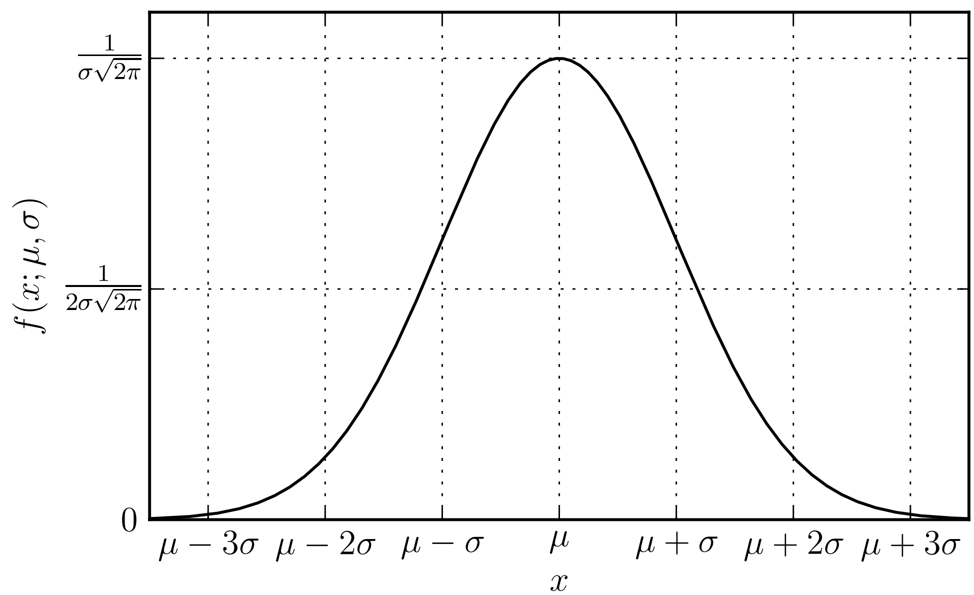
Thursday 13 October 2011

2 Normal Distribution

On the worksheet from last time, you considered one simple pdf for a continuous random variable, a uniform distribution. Now we're going to consider one which is more complicated, but very useful, the normal distribution, sometimes called the Gaussian distribution. The distribution depends on two parameters, μ and σ,[1] and if X follows a normal distribution with those parameters, we write $X \sim N(\mu, \sigma^2)$, and the pdf of X is

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{2.1}$$

The shape of the normal distribution is the classic "bell curve":



Some things we can notice about the normal distribution:

- It is non-zero for all x, so $x_{\min} = -\infty$ and $x_{\max} = \infty$
- The pdf is, in fact, normalized. This is kind of tricky to show, since it relies on the definite integral result

$$\int_{-\infty}^{\infty} e^{-z^2/2}\, dz = \sqrt{2\pi} \tag{2.2}$$

  This integral can't really be done by the usual means, since there is no ordinary function whose derivative is $e^{-z^2/2}$, and actually only the definite integral from $-\infty$ to ∞ (or from 0 to ∞) has a nice value. The proof of this identity is cute, but unfortunately it

[1] μ can be positive, negative or zero, but σ has to be positive

requires the use of a two-dimensional integral in polar coördinates, so it's beyond our scope at the moment.

- The parameter μ, as you might guess from the name, is the mean of the distribution. This is not so hard to see from the graph: the pdf is clearly symmetric about $x = \mu$, so that point should be both the mean and median of the distribution. It's also not so hard to show from the equations that
$$\int_{-\infty}^{\infty} x\, e^{-(x-\mu)^2/(2\sigma^2)}\, dx = \mu \int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)}\, dx \tag{2.3}$$
(and thus $E(X) = \mu$) by using the change of variables $y = x - \mu$.
- The parameter σ turns out to be the standard deviation. You can't really tell this from the graph, but it is possible to show that
$$V(X) = E((X-\mu)^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2\, e^{-(x-\mu)^2/(2\sigma^2)}\, dx \tag{2.4}$$
This is harder, though, since the integral
$$\int_{-\infty}^{\infty} (x-\mu)^2\, e^{-(x-\mu)^2/(2\sigma^2)}\, dx = \sigma^2 \int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)}\, dx \tag{2.5}$$
has to be done using integration by parts.
- Although the pdf depends on two parameters, neither one of them fundamentally changes the shape. Changing μ just slides the pdf back and forth. Changing σ just changes the scale of the horizontal axis. If we increase σ, the curve is stretched horizontally, and it has to be squeezed by the same factor vertically, in order to keep the area underneath it equal to 1 (since the pdf is normalized).

2.1 The Standard Normal Distribution

The fact that, for $X \sim N(\mu, \sigma^2)$, changing the parameters μ and σ just slides and stretches the x axis is the motivation for constructing a new random variable Z:
$$Z = \frac{X - \mu}{\sigma} \tag{2.6}$$
Now, because the expected values for continuous random variables still have the same behavior under linear transformations as for discrete rvs, i.e.,
$$E(aX + b) = aE(X) + b \tag{2.7a}$$
$$V(aX + b) = a^2 V(X) \tag{2.7b}$$
we can see that
$$E(Z) = \frac{E(X)}{\sigma} - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0 \tag{2.8a}$$
$$V(Z) = \frac{V(X)}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1 \tag{2.8b}$$

In fact, Z is itself normally distributed $Z \sim N(0,1)$. We call this special case of the normal distribution the *standard normal distribution*.

The standard normal distribution is often useful in calculating the probability that a normally distributed random variable falls in a certain finite interval:

$$P(a < X \leq b) = P\left(\frac{a-\mu}{\sigma} < Z \leq \frac{b-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-z^2/2}\, dz \tag{2.9}$$

The problem is that, as we noted before, there's no ordinary function whose derivative is $e^{-z^2/2}$. Still, the cumulative distribution function of a standard normal random variable is a perfectly sensible thing, and we could, for example, numerically integrate to find an approximate value for any z. So we define a function $\Phi(z)$ to be equal to that integral:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du \qquad \text{(definition)} \tag{2.10}$$

We can now write down the cdf for the original normal rv X:

$$F(x;\mu,\sigma) = P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{2.11}$$

and from the cdf we can find the probability for X to lie in any interval:

$$P(a < X \leq b) = F(b;\mu,\sigma) - F(a;\mu,\sigma) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{2.12}$$

2.1.1 Cumulative Distribution Function

The function $\Phi(z)$ has been defined so that it equals the cdf of Z:

$$F(z;1,0) = P(Z \leq z) = \int_{-\infty}^{z} f(u;0,1)\, du = \Phi(z) \tag{2.13}$$

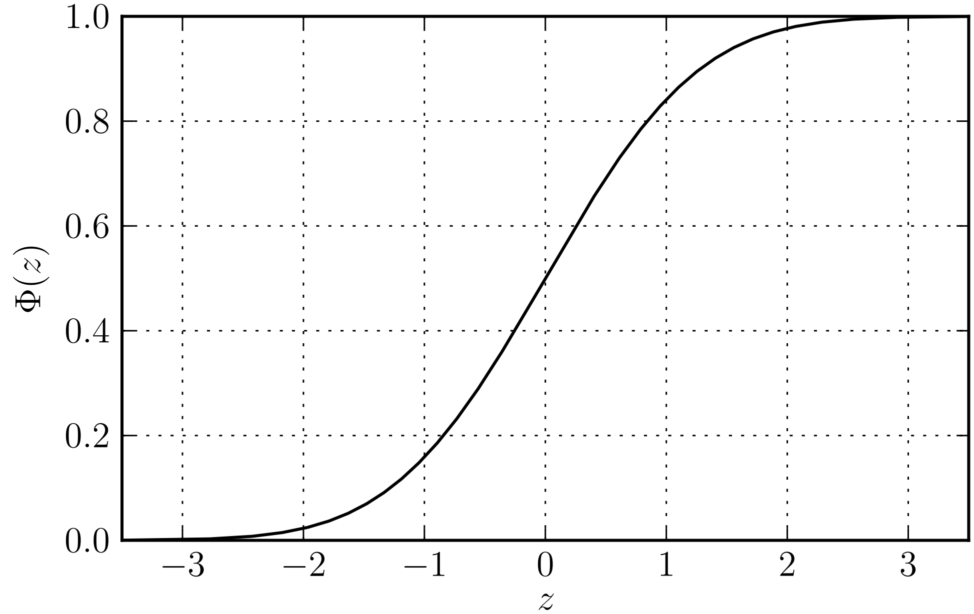
This identification of $\Phi(z)$ as a cdf tells us a few values exactly:

$$\Phi(-\infty) = 0 \tag{2.14a}$$

$$\Phi(0) = \frac{1}{2} \tag{2.14b}$$

$$\Phi(\infty) = 1 \tag{2.14c}$$

We know $\Phi(\infty)$ from the fact that the normal distribution is normalized. We can deduce $\Phi(0)$ from the fact that it's symmetric about its mean value, and therefore the mean value is also the median. For any other value of z, $\Phi(z)$ can in principle be calculated numerically. Here's what a plot of $\Phi(z)$ looks like:



Of course, since $\Phi(z)$ is an important function–like, for example, $\sin\theta$ or e^x–its value has been tabulated. Table A.3, in the back of the book, collects some of those values. Now, in the old days trigonometry books also had tables of values of trig functions, but now we don't need those because our calculators have sin and cos and e^x buttons. Unfortunately, our calculators don't have Φ buttons on them yet, so we have the table. Note that if you're using a mathematical computing environment like scipy or Mathematica or matlab which provides the "error function" $\mathrm{erf}(x)$, you can evaluate

$$\Phi(z) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)\right] \tag{2.15}$$

(That's how I made the plot of $\Phi(z)$.)

Some useful sample values of $\Phi(z)$ are

$$P(X \leq \mu + \sigma) = \Phi(1) \approx .8413 \tag{2.16a}$$
$$P(X \leq \mu + 2\sigma) = \Phi(2) \approx .9772 \tag{2.16b}$$
$$P(X \leq \mu + 3\sigma) = \Phi(3) \approx .9987 \tag{2.16c}$$
$$P(\mu - \sigma < X \leq \mu + \sigma) = \Phi(1) - \Phi(-1) \approx .6827 \tag{2.16d}$$
$$P(\mu - 2\sigma < X \leq \mu + 2\sigma) = \Phi(2) - \Phi(-2) \approx .9545 \tag{2.16e}$$
$$P(\mu - 3\sigma < X \leq \mu + 3\sigma) = \Phi(3) - \Phi(-3) \approx .9973 \tag{2.16f}$$

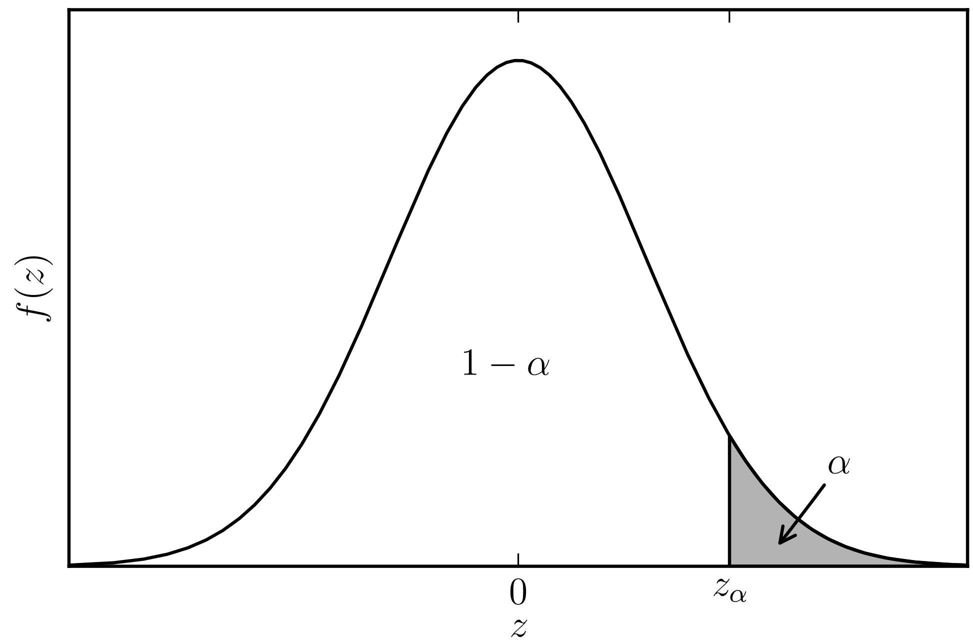
The last few in particular are useful to keep in mind: A given value drawn from a standard normal distribution has

- about a 2/3 (actually 68%) chance of being within one sigma of the mean
- about a 95% chance of being within two sigma of the mean
- about a 99.7% chance of being within three sigma of the mean

2.2 z_α Values and Percentiles

Often you want to go the other way, and ask, for example, what's the value that a normal random variable will only exceed 10% of the time. To get this, you basically need the inverse of the function $\Phi(z)$. The notation is to define z_α such that α (which can be between 0 and 1) is the probability that $Z > z_\alpha$:

$$P(Z > z_\alpha) = 1 - \Phi(z_\alpha) = \alpha \tag{2.17}$$



We know a few exact values:

$$z_0 = \infty \qquad z_{.5} = 0 \qquad z_1 = -\infty \tag{2.18}$$

The rest can be deduced from the table of $\Phi(z)$ values; a few of them are in Table 4.1 of Devore (on page 156 of the 8th edition). Note that his argument about $z_{.05}$ actually only shows that it's between 1.64 and 1.65. A more accurate value turns out to be $z_{.05} \approx 1.64485$ which means that, somewhat coïncidentally, $z_{.05} \approx 1.645$ is accurate to four significant figures. (But to three significant figures $z_{.05} \approx 1.64$.)

Note also that the definition of z-critical values is "backwards" from that of percentiles. E.g., $z_{.05}$ is the 95th percentile of the standard normal distribution.

2.3 Approximating Distributions of Discrete Random Variables

We described some aspects of a continuous random variable as a limiting case of a discrete random variable as the spacing between possible values went to zero. It turns out that continuous random variables, and normally-distributed ones in particular, can be good approximations of discrete random variables when the numbers involved are large. In this case, Δx doesn't go to zero, but it's small compared to the other numbers in the problem.

Suppose we have a discrete random variable X with pmf $p(x)$ which can take on values which are Δx apart.[2] The continuous random variable $\mathcal{X}$ with pdf $f(x)$ is a good approximation of X if

$$p(x) = P(X = x) \approx P(x - \Delta x/2 < \mathcal{X} < x + \Delta x/2) \approx f(x)\, \Delta x \tag{2.19}$$

One thing to watch out for is the discreteness of the original random variable X. For instance, if we want to estimate its cdf, $P(X \leq x)$, we should consider that the value $X = x$ corresponds to the range $x - \Delta x/2 < \mathcal{X} < x + \Delta x/2$, so

$$P(X \leq x) \approx P(\mathcal{X} < x + \Delta x/2) = \int_{-\infty}^{x+\Delta x/2} f(y)\, dy \tag{2.20}$$

Often the continuous random variable used in the approximation is a normally-distributed rv with the same mean and standard deviation as the original discrete rv. I.e., if $E(X) = \mu$ and $V(X) = \sigma^2$, it is often a good approximation to take

$$F(x) = P(X \leq x) \approx \int_{-\infty}^{x+\Delta x/2} f(y; \mu, \sigma)\, dy = \Phi\left(\frac{x + .5(\Delta x) - \mu}{\sigma}\right) \tag{2.21}$$

2.3.1 Example: Approximating the Binomial Distribution

As an example, consider a binomial random variable $X \sim \text{Bin}(n, p)$. Remember that for large values of n it was kind of a pain to find the cdf

$$P(X \leq x) = B(x; n, p) = \sum_{y=0}^{x} \binom{n}{y} p^y (1-p)^{n-y} = \sum_{y=0}^{x} \binom{n}{y} p^y q^{n-y} \tag{2.22}$$

(Did anyone forget what q is? I did. $q = 1 - p$.) But if we can approximate X as a normally-distributed random variable with $\mu = E(X) = np$ and $\sigma^2 = V(X) = np(1-p) = npq$, life is pretty simple (here $\Delta x = 1$ because x can take on integer values):

$$P(X \leq x) = B(x; n, p) \approx \Phi\left(\frac{x + .5 - np}{\sqrt{npq}}\right) \tag{2.23}$$

At the very least, we only have to look in one table, rather than in a two-parameter family of tables!

Practice Problems

4.29, 4.31, 4.35, 4.47, 4.53, 4.55

[2] This is made even simpler if the discrete rv can take on integer values, so that $\Delta x = 1$, but it also works for other examples, like SAT scores, for which $\Delta x = 10$.

Tuesday 18 October 2011

3 Other Continuous Distributions

Sections 4.4 and 4.5 unleash upon us a dizzying array of other distribution functions. (I count six or seven.) My goal here is to put these into context so we can understand how they fit together. First, a qualitative overview:

- The *exponential distribution* occurs in the context of a Poisson process. It is the pdf of the time we have to wait for a Poisson event to occur.
  - The *gamma* and *Weibull distributions* are variants of the exponential distribution with a modified shape. They describe physical situations where the expected waiting time changes depending on how long we've been waiting.
- The *chi-squared distribution* is the pdf of the sum of the squares of one or more independent standard normal random variables. This pdf turns out to be a special case of the gamma distribution.
- The *log-normal distribution* is the pdf of a rv whose logarithm obeys the normal distribution.
- The *beta distribution* is a type of pdf for a quantity which can be defined over a finite interval.

Part of the complexity of the formulas for these various random variables arises from several things which are not essential to the fundamental shape of the distribution. Consider a normal distribution

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{3.1}$$

First, notice that the factor of $\frac{1}{\sigma\sqrt{2\pi}}$ does not depend on the value x of the random variable X. It's a constant (which happens to depend on the parameter σ) which is just needed to normalize the distribution. So we can write

$$f(x;\mu,\sigma) = \mathcal{N}(\sigma) e^{-(x-\mu)^2/(2\sigma^2)} \tag{3.2}$$

or

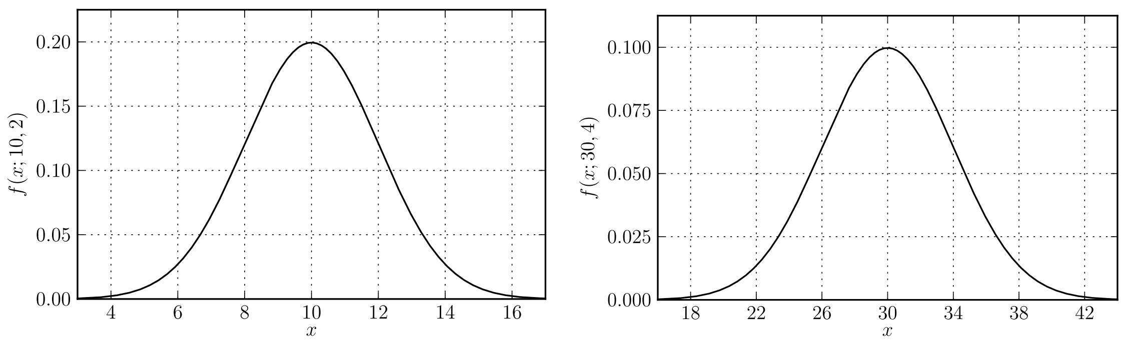
$$f(x;\mu,\sigma) \propto e^{-(x-\mu)^2/(2\sigma^2)} \tag{3.3}$$

We will try to separate the often complicated forms of the normalization constants from the often simpler forms of the x-dependent parts of the distributions.

Also, notice that the parameters μ and σ affect the location and scale of the pdf, but not the shape, and if we define the rv $Z = \frac{X-\mu}{\sigma}$, the pdf is

$$f(z) \propto e^{-z^2/2} \tag{3.4}$$

BTW, what I mean by "doesn't change the shape" is that if I plot a normal pdf with different parameters, I can always change the ranges of the axes to make it look like any other one. For example:



3.1 Exponential Distribution

Suppose we have a Poisson process with rate λ, so that the probability mass function of the number of events occurring in any period of time of duration t is

$$P(n \text{ events in } t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \tag{3.5}$$

In particular, evaluating the pmf at $n = 0$ we get

$$P(\text{no events in } t) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t} \tag{3.6}$$

Now pick some arbitrary moment and let X be the random variable describing how long we have to wait for the next event. If $X \leq t$ that means there are one or more events occurring in the interval of length t starting at that moment, so the probability of this is

$$P(X \leq t) = 1 - P(\text{no events in } t) = 1 - e^{-\lambda t} \tag{3.7}$$

But that is the definition of the cumulative distribution function, so

$$F(x; \lambda) = P(X \leq x) = 1 - e^{-\lambda x} \tag{3.8}$$

Note that it doesn't make sense for X to take on negative values, which is good, since $F(0) = 0$. This means that technically,

$$F(x; \lambda) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases} \tag{3.9}$$

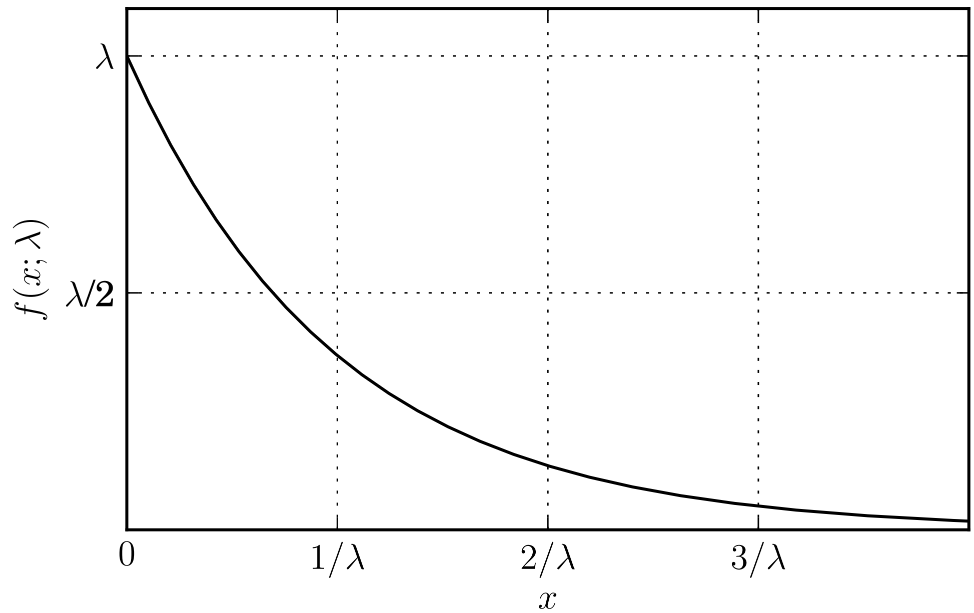
We can differentiate (with respect to x) to get the pdf

$$f(x; \lambda) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases} \tag{3.10}$$

If we isolate the x dependence, we can write the slightly simpler

$$f(x; \lambda) \propto e^{-\lambda x} \qquad x \geq 0 \tag{3.11}$$

In any event, we see that changing λ just changes the scale of the x variable, so we can draw one shape:



You can show, by using integration by parts to do the integrals, that

$$E(X) = \int_0^\infty x\,\lambda\,e^{-\lambda x}\,dx = \frac{1}{\lambda} \tag{3.12}$$

and

$$E(X^2) = \int_0^\infty x^2\,\lambda\,e^{-\lambda x}\,dx = \frac{2}{\lambda^2} \tag{3.13}$$

so

$$V(X) = E(X^2) - [E(X)]^2 = \frac{1}{\lambda^2} \tag{3.14}$$

3.1.1 Gamma and Weibull Distributions

Sometimes you also see the parameter in an exponential distribution written as $\beta = 1/\lambda$ rather than λ, so that

$$f(x;\beta) \propto e^{-x/\beta} \qquad x \geq 0 \tag{3.15}$$

The *gamma distribution* is a generalization which adds an additional parameter $\alpha > 0$, so that

$$f(x;\alpha,\beta) \propto \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta} \qquad x \geq 0 \qquad \text{gamma distribution} \tag{3.16}$$

The *Weibull distribution* uses the α to influence the shape in a slightly different way

$$f(x;\alpha,\beta) \propto \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-(x/\beta)^\alpha} \qquad x \geq 0 \qquad \text{Weibull distribution} \tag{3.17}$$

In either case, α (which need not be an integer) really does influence the shape of the distribution, and $\alpha = 1$ reduces to the exponential distribution.

3.1.2 The Gamma Function

Most of the complication associated with the gamma distribution in particular is associated with finding the normalization constant. If we write

$$f(x;\alpha,\beta) = \mathcal{N}(\alpha,\beta)\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta} \tag{3.18}$$

then

$$\frac{1}{\mathcal{N}(\alpha,\beta)} = \int_0^\infty \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta}\,dx = \beta\int_0^\infty u^{\alpha-1}e^{-u}\,du = \frac{\beta}{\mathcal{N}(\alpha,1)} \tag{3.19}$$

This integral

$$\frac{1}{\mathcal{N}(\alpha,1)} = \int_0^\infty x^{\alpha-1}e^{-x}\,dx = \Gamma(\alpha) \tag{3.20}$$

is defined as the "Gamma function", which has lots of nifty properties:

- $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ if $\alpha > 1$ (which can be shown by integration by parts)
- $\Gamma(1) = \int_0^\infty e^{-x}\,dx = 1$ from which it follows that $\Gamma(n) = (n-1)!$ if n is a positive integer
- $\Gamma(1/2) = \sqrt{\pi}$ which actually follows from $\int_{-\infty}^\infty e^{-z^2/2}\,dz = \sqrt{2\pi}$

3.2 Chi-Squared Distribution

Consider the square of a standard normal random variable, $X = Z^2$. Its cdf is

$$F(x) = P(X \le x) = P(Z^2 \le x) = P(-\sqrt{x} \le Z \le \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) \tag{3.21}$$

which we can differentiate to get the pdf

$$f(x) = F'(x) = \frac{dP}{dz}\frac{dz}{dx} = \left(\frac{2}{\sqrt{2\pi}}e^{-x/2}\right)\left(\frac{1}{2\sqrt{x}}\right) \propto x^{-1/2}e^{-x/2} \tag{3.22}$$

This is a special case of chi-squared distribution, with one degree of freedom. It turns out that if you take the sum of the squares of ν independent standard normal random variables, $X = \sum_{k=1}^{\nu}(Z_\nu)^2$, it obeys a chi-squared distribution with ν degrees of freedom,

$$f(x) \propto x^{(\nu/2)-1}e^{-x/2} \tag{3.23}$$

But if we compare this to equation (3.16) we see that the chi-squared distribution with ν degrees of freedom is just a Gamma distribution with $\alpha = \nu/2$ and $\beta = 2$.

Practice Problems

4.59, 4.61, 4,67, 4.69, 4.71, 4.79, 4.85

Thursday 20 October 2011

3.3 Log-Normal Distribution

While many random variables can be approximated as obeying a normal distribution, there are some situations where this doesn't make sense. Some variables are more naturally associated with multiplication than addition. For instance, in the early days of astronomy, distance measurements of distant object might only be known to about a factor of two. That would mean that an object at distance d might be reported as having a distance of $d + d = 2d$ or greater $1/6$ ($\approx 1 - \Phi(1)$) of the time. If the reported distance were governed by a normal distribution with a mean of d and a standard deviation of d, that would mean that we'd also expect to measure a distance of $d - d = 0$ or less $1/6$ ($\approx \Phi(-1)$) of the time. But that clearly makes no sense, since a distance measurement is not going to be negative. Instead, we would like some sort of a distribution where $1/6$ of the time the measurement will be too small by a factor of two, i.e., be less than $d/2$.

Generalizing this example, replacing d with one constant A and the factor of d with another factor B, we can consider a random variable X for which

$$P(A/C \le X \le A) = P(A \le X \le AC) \qquad \text{for any } C \tag{3.24a}$$
$$P(A/B \le X \le AB) \approx .68 \tag{3.24b}$$
$$P(A/B^2 \le X \le AB^2) \approx .95 \tag{3.24c}$$
$$P(A/B^3 \le X \le AB^3) \approx .997 \tag{3.24d}$$

This behaves similarly to a normally-distributed rv, except that multiplication has replaced addition. If we instead talk about the natural logarithm $\ln X$, it obeys

$$P(\ln A - \ln C \le \ln X \le \ln A) = P(\ln A \le \ln X \le \ln A + \ln C) \qquad \text{for any } C \tag{3.25a}$$
$$P(\ln A - \ln B \le \ln X \le \ln A + \ln B) \approx .68 \tag{3.25b}$$
$$P(\ln A - 2\ln B \le \ln X \le \ln A + 2\ln B) \approx .95 \tag{3.25c}$$
$$P(\ln A - 3\ln B \le \ln X \le \ln A + 3\ln B) \approx .997 \tag{3.25d}$$

But those are just the properties we get if $\ln X$ is a normally-distributed random variable with $\mu = \ln A$ and $\sigma = \ln B$. This allows us to deduce the cdf $F(x)$ and from it the pdf $f(x) = F'(x)$. For convenience we define a the random variable

$$Z = \frac{\ln X - \mu}{\sigma} \tag{3.26}$$

which will obey a standard normal distribution, and also the function

$$z(x) = \frac{\ln x - \mu}{\sigma} \tag{3.27}$$

Then

$$F(x) = P(X \le x) = \Phi\left(z(x)\right) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right) \tag{3.28}$$

We can differentiate this using the chain rule:

$$f(x) = F'(x) = \frac{dF}{dx} = \frac{d\Phi}{dz}\frac{dz}{dx} = \Phi'(z)\, z'(x) \tag{3.29}$$

The derivative of $\Phi(z)$ with respect to z is defined to be

$$\Phi'(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2} = \frac{1}{\sqrt{2\pi}}\, e^{-(\ln x - \mu)^2/(2\sigma^2)} \tag{3.30}$$

and the derivative of $z(x)$ with respect to x is

$$z'(x) = \frac{d}{dx}\frac{\ln x - \mu}{\sigma} = \frac{1}{\sigma}\frac{d}{dx}\ln x = \frac{1}{\sigma}\frac{1}{x} \tag{3.31}$$

So

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,\frac{1}{x}\, e^{-(\ln x - \mu)^2/(2\sigma^2)} \tag{3.32}$$

3.4 Beta Distribution

This is a family of distributions used when X is confined to a finite range of values, from A to B. First off, we can rescale to define

$$Y = \frac{X - A}{B - A} \tag{3.33}$$

so that Y can fall between 0 and 1; then

$$1 - Y = \frac{B - X}{B - A} \tag{3.34}$$

The distribution uses positive parameters α and β and is defined with

$$f(y) \propto y^{\alpha-1}(1-y)^{\beta-1} \qquad 0 < y < 1 \tag{3.35}$$

i.e.,

$$f(x) \propto \left(\frac{x - A}{B - A}\right)^{\alpha-1} \left(\frac{B - x}{B - A}\right)^{\beta-1} \qquad A < x < B \tag{3.36}$$

Note that the special case of $\alpha = 1$, $\beta = 1$ is the uniform distribution.

Again, the extra complications (see the corresponding form of the pdf in Devore) arise from getting the normalization constant right.

3.5 Worksheet

- worksheet
- solutions

4 Probability Plots

Back in Chapter One, we defined a bunch of properties of a sample: mean, variance, standard deviation, median, and various percentiles. We also defined the corresponding properties for the underlying population. Since then, we've been considering random variables, which can be thought of as based on conceptual populations, and defined analogous properties of mean, variance, standard deviation, median and percentiles associated with the underlying probability distribution.

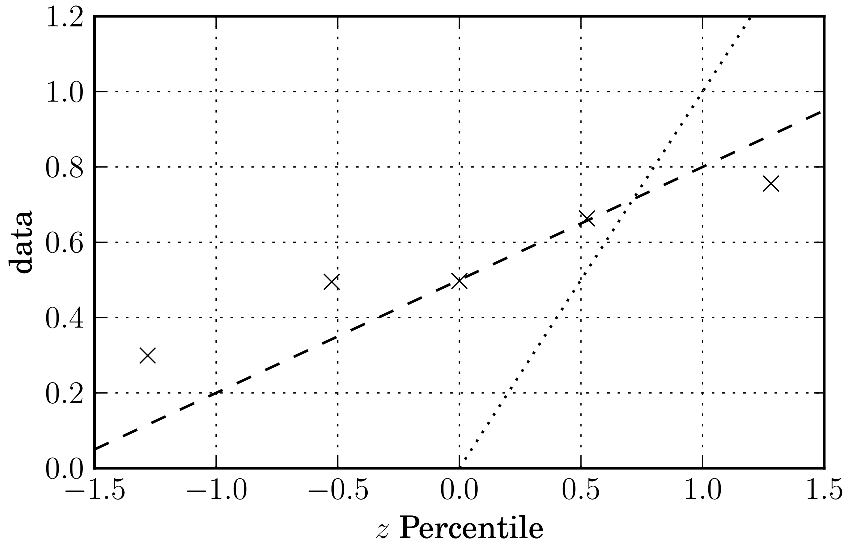
One thing we can do is consider a data sample, and check how likely it is to have originated from a given probability distribution. (We'll only do this qualitatively at this point.) One simple thing to do is compare the median: is the sample median close to the median of the proposed probability distribution? We can also start to ask this about various percentiles of the data: do, for example, the top 10% of the data lie above the 90th percentile of the probability distribution? But which percentiles do we check? We can't really check more percentiles than we have data points, and we'll let the points themselves tell us where to check. Recall that if we have an odd number of points, the middle one (once they're sorted into order) is the median, but if we have an even number, we have to interpolate. The idea, if we have e.g., 5 points, is that two and a half lie below and two and a half lie above the middle value. We can do the same trick with the other points: the lowest value has half a point below and 4.5 above it, so it's the 10th percentile of the sample. With 5 points we get the following correspondence:

Sample #	1	2	3	4	5
Pct below	10	30	50	70	90

(In general, the percentile corresponding to point i after sorting a sample of n points is $100(i - .5)/n$.) We plot each value against the corresponding percentile of the distribution we want, so if we have five samples the lowest value in the sample is plotted against the 10th percentile of the distribution, the next lowest against the 30th percentile, etc. Let's look at this for an example data set, and see if it fits a standard normal distribution, from which we can use z_α with $\alpha = 1 - (i - .5)/n$.

Sample #	1	2	3	4	5
Percentage	10	30	50	70	90
z percentile	-1.282	-0.524	0.000	0.524	1.282
Sample obs.	0.299	0.495	0.497	0.663	0.756

The data points don't agree at all well with the corresponding percentiles of the standard normal distribution, which we can see by drawing a dotted line with $y = x$ on the corresponding plot:



Now, this is not so surprising, since I generated the data using a normal distribution with $\mu = .3$ and $\sigma = .5$. If we work out the percentiles of the distribution with those parameters, the values are much closer:

Sample #	1	2	3	4	5
Percentage	10	30	50	70	90
z percentile	-1.282	-0.524	0.000	0.524	1.282
Sample obs.	0.299	0.495	0.497	0.663	0.756
$N(0.5, 0.3^2)$ *percentile*	0.116	0.343	0.500	0.657	0.884

Now, we may often want to check whether data are consistent with a normal distribution without specifying μ and σ for that distribution. The cool thing is that we don't have to, because any normal random variable $X \sim N(\mu, \sigma^2)$ is related to a corresponding standard normal random variable by

$$Z = \frac{X - \mu}{\sigma} \tag{4.1}$$

or

$$X = \mu + \sigma \cdot Z \tag{4.2}$$

This means that the $100(1 - \alpha)$ percentile x_α is

$$x_\alpha = \mu + \sigma \cdot z_\alpha \tag{4.3}$$

and a sample is consistent with a normal distribution if it lies close to a straight line on a normal probability plot. The dashed line on the plot above is $.5 + .3z$, which passes through the appropriate percentiles for a normal distribution with $\mu = .5$ and $\sigma = .3$.

Practice Problems

4.87, 4.89, 4.91, 4.93, 4.97, 4.111