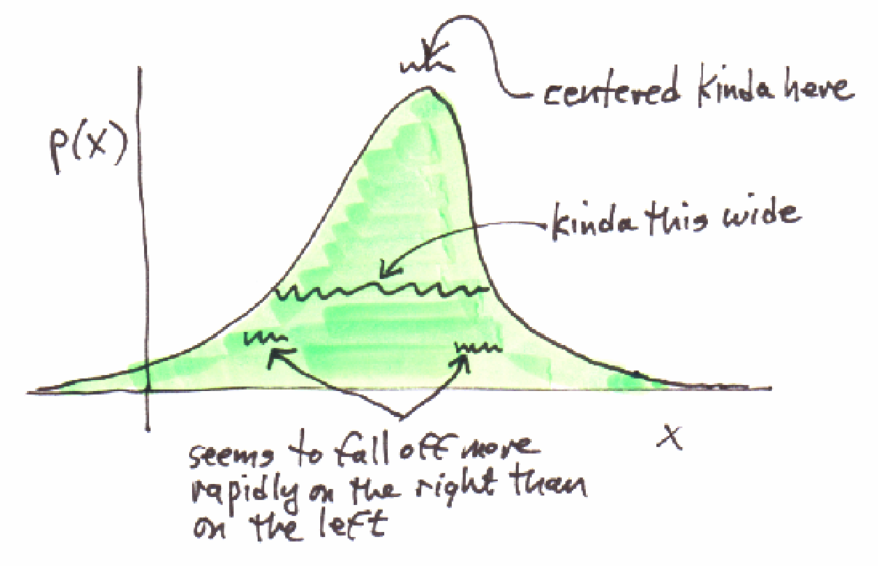
CS395T
Computational Statistics with Application to Bioinformatics

Prof. William H. Press
Spring Term, 2009
The University of Texas at Austin

Unit 3: Common Distributions

Many, though not all, common distributions look sort-of like this:



We already saw the beta distribution with $\alpha, \beta > 0$ as an example on the interval [0,1], and the Towne family example (not any simple function). We'll see more examples soon.

Suppose we want to <u>summarize</u> *p(x)* by a single number *a*, its "center". Let's find the value *a* that minimizes the mean-square distance of the "typical" value *x*:

expectation notation:
$\langle \text{anything} \rangle \equiv \int_x (\text{anything})\, p(x) dx$
expectation is linear, etc.

$$\text{minimize:} \quad \Delta^2 \equiv \left\langle (x-a)^2 \right\rangle = \left\langle x^2 - 2ax + a^2 \right\rangle$$
$$= \left(\left\langle x^2 \right\rangle - \langle x \rangle^2 \right) + \left(\langle x \rangle - a \right)^2$$

This is the variance Var(x), but all we care about here is that it doesn't depend on a.

(in physics this is called the "parallel axis theorem")

The minimum is obviously $a = \langle x \rangle$. (Take derivative wrt a and set to zero if you like mechanical calculations.)

Why mean-square? Why not mean-absolute? Try it!

$$\Delta = \langle |x-a| \rangle = \int_{-\infty}^{\infty} |x-a|\, p(x) dx$$

$$= \int_{-\infty}^{a} (a-x)\, p(x) dx + \int_{a}^{\infty} (x-a)\, p(x) dx$$

So,

$$0 = \frac{d\Delta}{da} = \int_{-\infty}^{a} p(x) dx + 0 - \int_{a}^{\infty} p(x) dx + 0$$

Integrand at a

$$\Rightarrow$$

$$\int_{-\infty}^{a} p(x) dx = \int_{a}^{\infty} p(x) dx = \tfrac{1}{2}$$

$\Rightarrow a$ is the median value

Mean and median are both “measures of central tendency”.

Higher moments, centered moments defined by

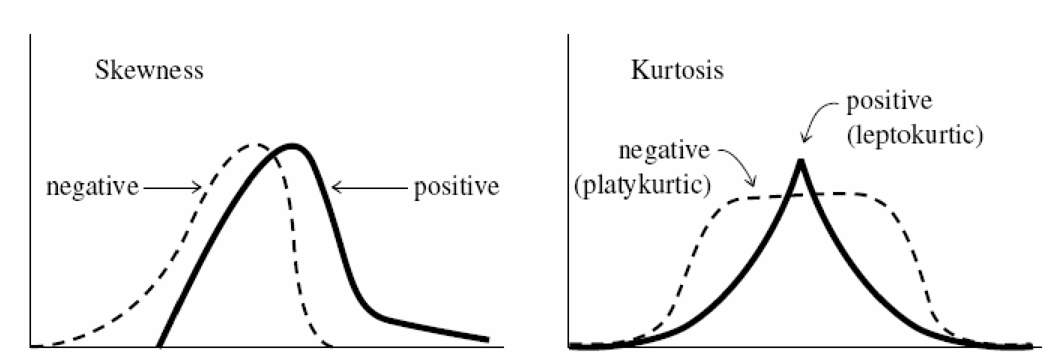
$$\mu_i \equiv \langle x^i \rangle = \int x^i \, p(x) dx$$
$$M_i \equiv \left\langle (x - \langle x \rangle)^i \right\rangle = \int (x - \langle x \rangle)^i \, p(x) dx$$

The centered second moment M_2, the variance, is by far most useful

$$M_2 \equiv \text{Var}(x) \equiv \left\langle (x - \langle x \rangle)^2 \right\rangle = \left\langle x^2 \right\rangle - \langle x \rangle^2$$
$$\sigma(x) \equiv \sqrt{\text{Var}(x)}$$

"standard deviation" summarizes a distribution's half-width (r.m.s. deviation from the mean)

Third and fourth moments also have "names"



But generally wise to be cautious about using high moments. Otherwise perfectly good distributions don't have them at all (divergent). And (related) it can take a lot of data to measure them accurately.

Mean and variance are additive over independent random variables:

$$\overline{(x+y)} = \overline{x} + \overline{y} \qquad \mathrm{Var}(x+y) = \mathrm{Var}(x) + \mathrm{Var}(x)$$

note “bar” notation, equivalent to < >

Certain combinations of higher moments are also additive. These are called semi-invariants.

$$I_2 = M_2 \qquad I_3 = M_3 \qquad I_4 = M_4 - 3M_2^2$$

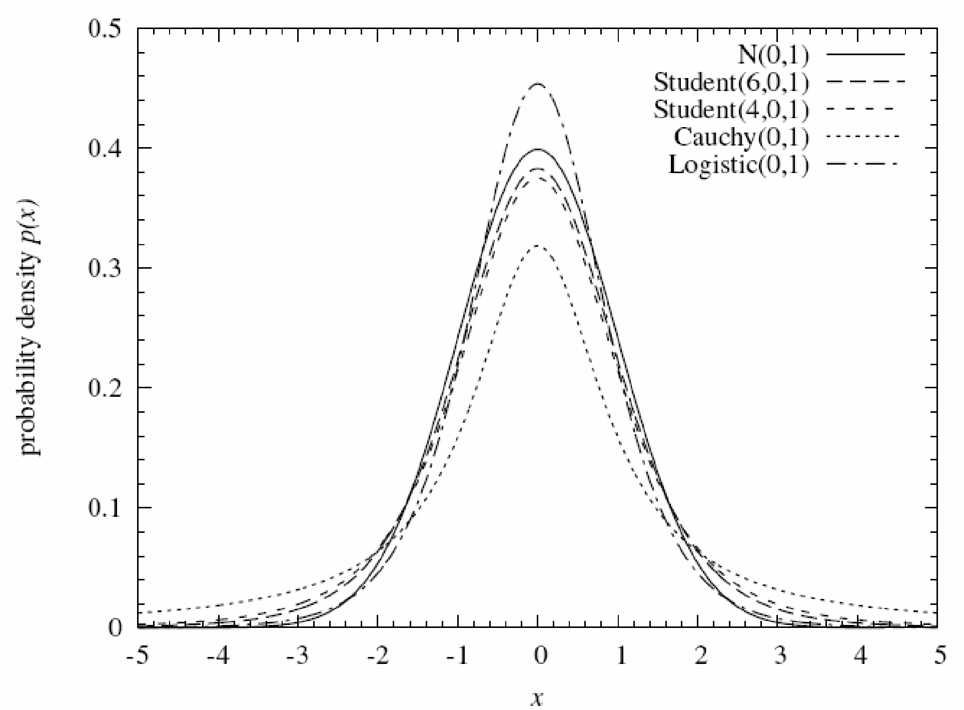
$$I_5 = M_5 - 10M_2M_3 \qquad I_6 = M_6 - 15M_2M_4 - 10M_3^2 + 30M_2^3$$

Skew and kurtosis are dimensionless combinations of semi-invariants

$$\mathrm{Skew}(x) = I_3/I_2^{3/2} \qquad \mathrm{Kurt}(x) = I_4/I_2^2$$

A Gaussian has all of its semi-invariants higher than I_2 equal to zero.
A Poisson distribution has all of its semi-invariants equal to its mean.

This is a good time to review some standard (i.e., frequently occurring) distributions:



Normal (Gaussian):

$$x \sim \mathrm{N}(\mu, \sigma), \qquad \sigma > 0$$

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right)$$

Student:

$$t \sim \text{Student}(\nu, \mu, \sigma), \qquad \nu > 0, \ \sigma > 0$$

$$p(t) = \frac{\Gamma(\frac{1}{2}[\nu+1])}{\Gamma(\frac{1}{2}\nu)\sqrt{\nu\pi}\sigma}\left(1 + \frac{1}{\nu}\left[\frac{t-\mu}{\sigma}\right]^2\right)^{-\frac{1}{2}(\nu+1)}$$

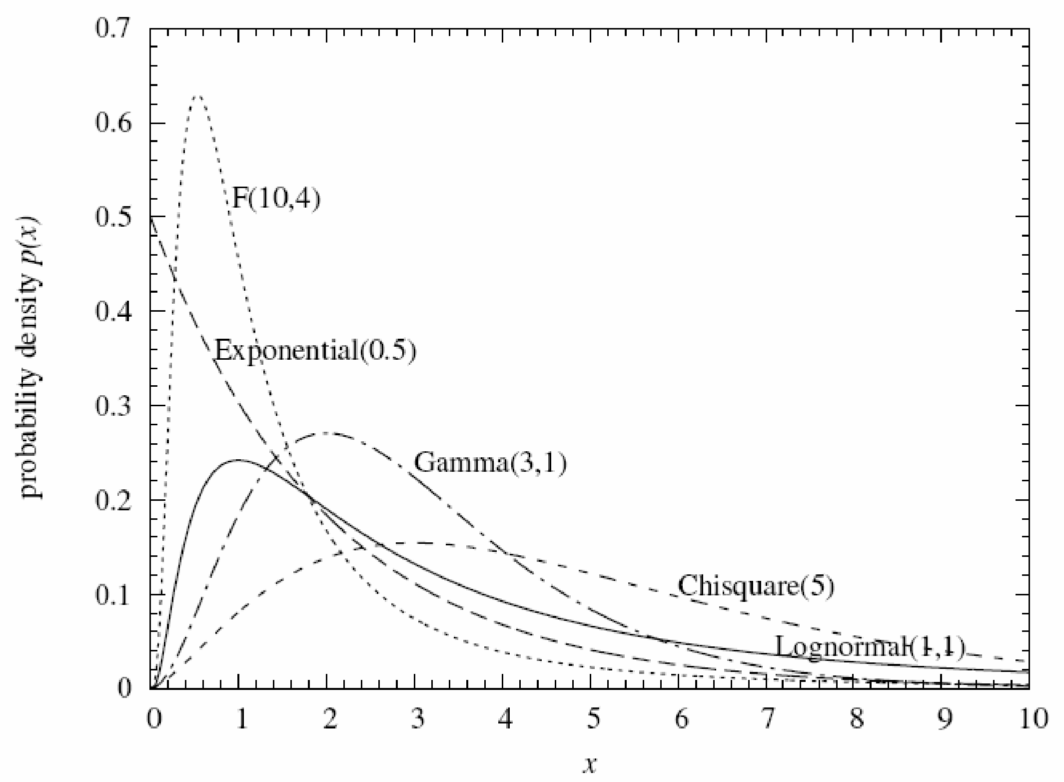
"bell shaped" but you get to specify the power with which the tails fall off. Normal and Cauchy are limiting cases. (Also occurs in some statistical tests.)

note that σ is not (quite) the standard deviation!

we'll see uses for "heavy-tailed" distributions later

"Student" was actually William Sealy Gosset (1876-1937), who spent his entire career at the Guinness brewery in Dublin, where he rose to become the company's Master Brewer.

Common distributions on positive real line:



Exponential:

$$x \sim \text{Exponential}(\beta), \qquad \beta > 0$$
$$p(x) = \beta \exp(-\beta x), \qquad x > 0$$

Lognormal:

$$x \sim \text{Lognormal}(\mu, \sigma), \qquad \sigma > 0$$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2}\left[\frac{\log(x) - \mu}{\sigma}\right]^2\right), \qquad x > 0 \tag{6.14.31}$$

Note the required extra factor of x^{-1} in front of the exponential: The density that is "normal" is $p(\log x) d \log x$.

While μ and σ are the mean and standard deviation in $\log x$ space, they are *not* so in x space. Rather,

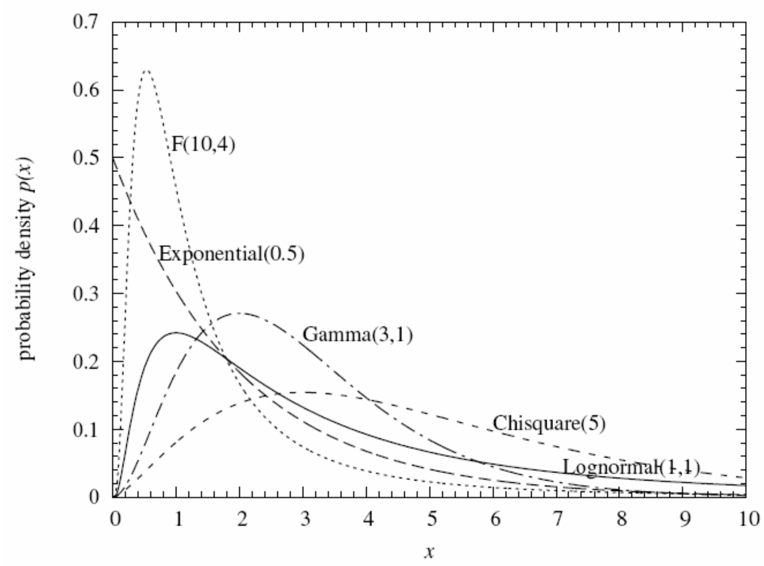
$$\begin{aligned} \text{Mean}\{\text{Lognormal}(\mu, \sigma)\} &= e^{\mu + \frac{1}{2}\sigma^2} \\ \text{Var}\{\text{Lognormal}(\mu, \sigma)\} &= e^{2\mu} e^{\sigma^2}(e^{\sigma^2} - 1) \end{aligned} \tag{6.14.32}$$

```
syms mu sig positive
syms x
Pi = sym(pi);
p = (1/sqrt(2*Pi))*(1/(sig*x))*exp(-(1/2)*(log(x)-mu)^2/sig^2)
p =
1/2*2^(1/2)/pi^(1/2)/sig/x*exp(-1/2*(log(x)-mu)^2/sig^2)
norm = int(p,0,inf)
norm =
1
mean = int(x*p,0,inf)
mean =
exp(mu+1/2*sig^2)
sig = simplify(sqrt(int(x^2*p,0,inf)-mean^2))
sig =
exp(mu+1/2*sig^2)*(exp(sig^2)-1)^(1/2)
```

Gamma distribution:

$$x \sim \text{Gamma}(\alpha, \beta), \qquad \alpha > 0,\ \beta > 0$$

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0$$



$$\text{Mean}\{\text{Gamma}(\alpha, \beta)\} = \alpha/\beta$$

$$\text{Var}\{\text{Gamma}(\alpha, \beta)\} = \alpha/\beta^2$$

When $\alpha \geq 1$ there is a single mode at $x = (\alpha - 1)/\beta$

- Gamma and Lognormal are both commonly used as convenient 2-parameter fitting functions for “peak with tail” positive distributions.
- Both have parameters for peak location and width.
- Neither has a separate parameter for how the tail decays.
  - Gamma: exponential decay
  - Lognormal: long-tailed (exponential of square of log)

Chi-square distribution (we'll use this a lot!)

Has only one parameter ν that determines both peak location and width.
ν is often an integer, called "number of degrees of freedom" or "DF"

the independent variable is χ^2, not χ

$$\chi^2 \sim \text{Chisquare}(\nu), \qquad \nu > 0$$

$$p(\chi^2)d\chi^2 = \frac{1}{2^{\frac{1}{2}\nu}\Gamma(\frac{1}{2}\nu)}(\chi^2)^{\frac{1}{2}\nu-1}\exp\left(-\frac{1}{2}\chi^2\right)d\chi^2, \qquad \chi^2 > 0$$

$$\text{Mean}\{\text{Chisquare}(\nu)\} = \nu$$
$$\text{Var}\{\text{Chisquare}(\nu)\} = 2\nu$$

When $\nu \geq 2$ there is a single mode at $\chi^2 = \nu - 2$

It's actually just a special case of Gamma, namely Gamma(ν/2,1/2)

For univariate distributions, one wants efficient computational methods for all of:

- PDF *p(x)*
- CDF *P(x)*
- Inverse of CDF *x(P)*
- Random deviates drawn from it (we'll get to soon)

NR3 has classes for many common distributions, with algorithms for p, cdf, and inverse cdf.

```
struct Normaldist : Erf {
Normal distribution, derived from the error function Erf.
    Doub mu, sig;
    Normaldist(Doub mmu = 0., Doub ssig = 1.) : mu(mmu), sig(ssig) {
    Constructor. Initialize with μ and σ. The default with no arguments is N(0, 1).
        if (sig <= 0.) throw("bad sig in Normaldist");
    }
    Doub p(Doub x) {
    Return probability density function.
        return (0.398942280401432678/sig)*exp(-0.5*SQR((x-mu)/sig));
    }
    Doub cdf(Doub x) {
    Return cumulative distribution function.
        return 0.5*erfc(-0.707106781186547524*(x-mu)/sig);
    }
    Doub invcdf(Doub p) {
    Return inverse cumulative distribution function.
        if (p <= 0. || p >= 1.) throw("bad p in Normaldist");
        return -1.41421356237309505*sig*inverfc(2.*p)+mu;
    }
};
```

Matlab and Mathematica both have many distributions, e.g.,

```
chi2pdf(x,v)
chi2cdf(x,v)
chi2inv(p,v)
```