

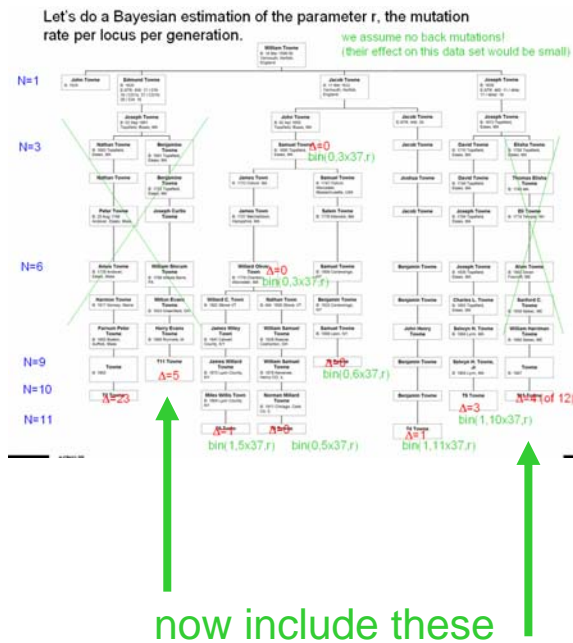
**CS395T**  
**Computational Statistics with**  
**Application to Bioinformatics**

Prof. William H. Press  
Spring Term, 2011  
The University of Texas at Austin

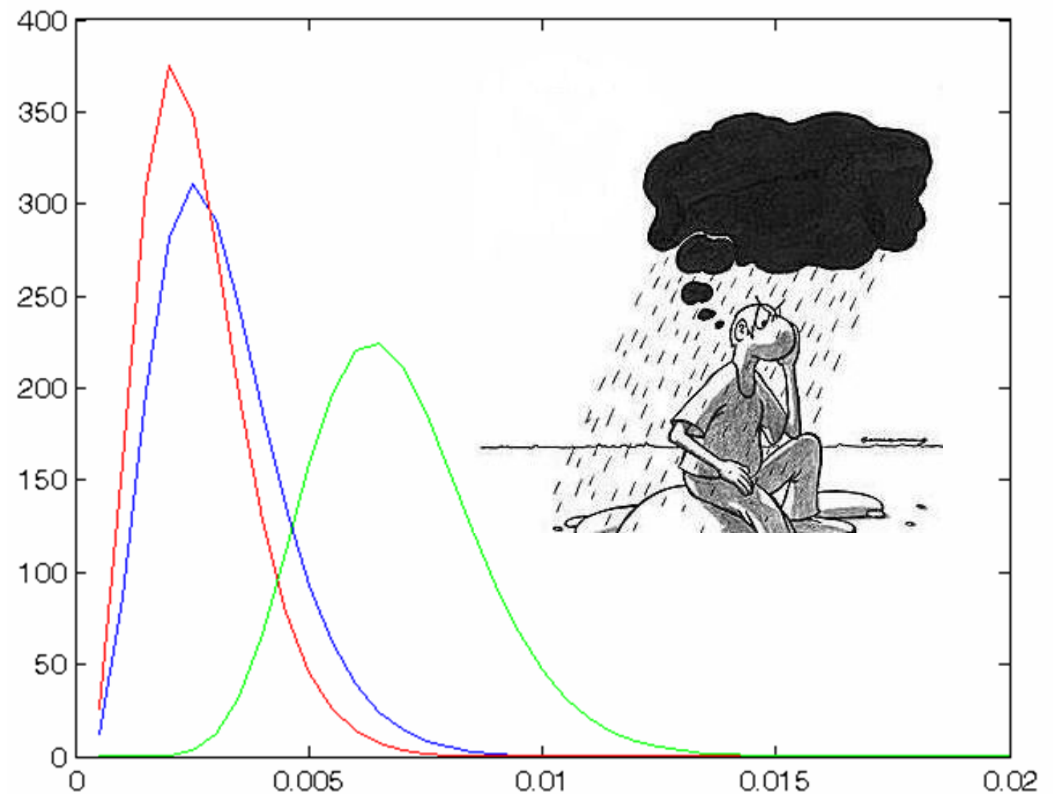
Lecture 4

A small cloud: The way we “trimmed” the data mattered. (And should trouble us a bit!) Here’s the effect of including T11 and T13, both of which seemed to be outliers:

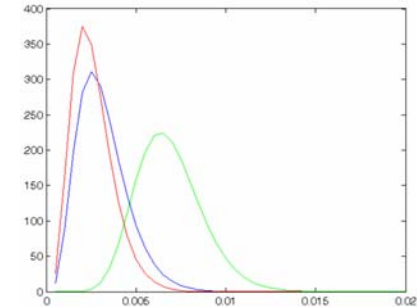
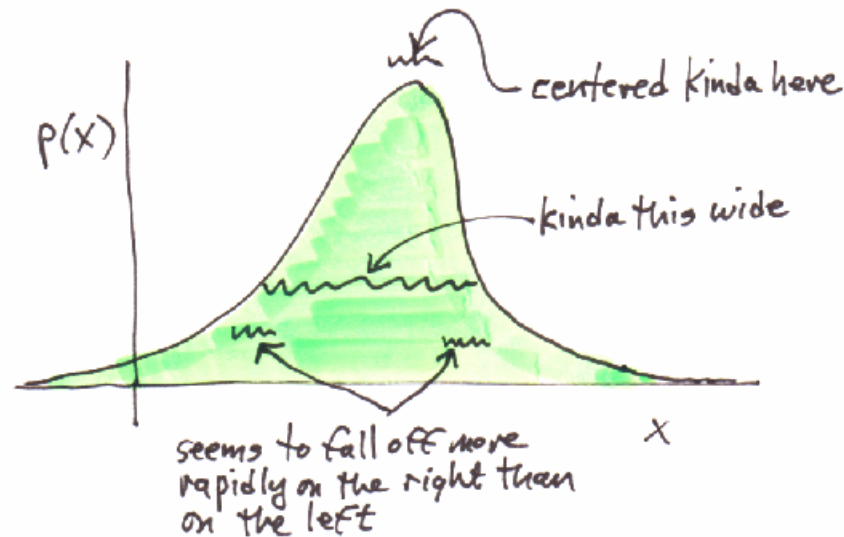
$$P(\text{data}|r) = [\text{old model}] \times \text{bin}(5, 9 \times 37, r) \text{bin}(4, 10 \times 12, r)$$



Editing outliers is a tricky issue that we will return to when we learn about mixture models.



We are often interested in distributions that have some kind of localization (because why would we be interested if they didn't?)



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 summarize  $p(x)$  by a single number  $a$ , its "value". Let's find the value  $a$  that minimizes the mean-square discrepancy of the "typical" value  $x$ :


Recall expectation notation:

$$\langle \text{anything} \rangle \equiv \int_x (\text{anything}) p(x) dx$$

i.e., the weighted average of “anything”, weighted by the probable values of  $x$ .  
Expectation is linear over “anything” (sums, constants times, etc.).

$$\begin{aligned} \text{minimize: } \Delta^2 &\equiv \langle (x - a)^2 \rangle = \langle x^2 - 2ax + a^2 \rangle \\ &= (\langle x^2 \rangle - \langle x \rangle^2) + (\langle x \rangle - a)^2 \end{aligned}$$

This is the variance  $\text{Var}(x)$ ,  
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!

$$\begin{aligned}\Delta &= \langle |x - a| \rangle = \int_{-\infty}^{\infty} |x - a| p(x) dx \\ &= \int_{-\infty}^a (a - x) p(x) dx + \int_a^{\infty} (x - a) p(x) dx\end{aligned}$$

So,

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

$\Rightarrow$

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

$\Rightarrow a$  is the median value

Integrand at  $a$



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

Higher moments, centered moments are conventionally defined by

$$\mu_i \equiv \langle x^i \rangle = \int x^i p(x) dx$$

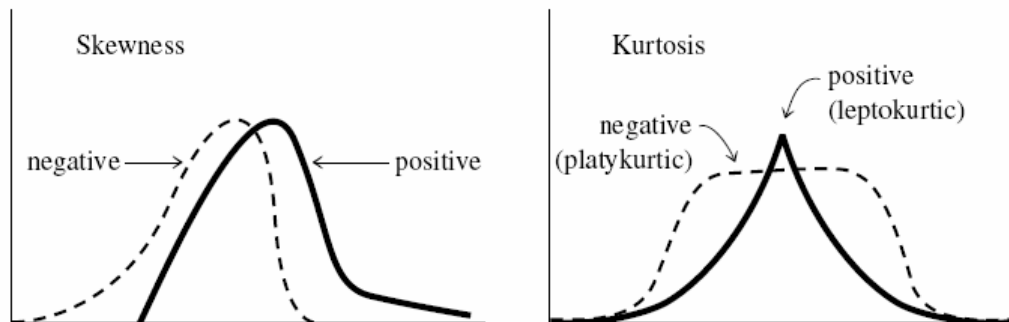
$$M_i \equiv \langle (x - \langle x \rangle)^i \rangle = \int (x - \langle x \rangle)^i p(x) dx$$

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

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

$$\sigma(x) \equiv \sqrt{\text{Var}(x)} \leftarrow \text{“standard deviation” summarizes a distribution’s half-width (r.m.s. deviation from the mean)}$$

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:

$$\overline{(x + y)} = \bar{x} + \bar{y} \quad \text{Var}(x + y) = \text{Var}(x) + \text{Var}(y)$$

note "bar" notation, equivalent to  $\langle \rangle$

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

$$\begin{aligned} I_2 &= M_2 & I_3 &= M_3 & I_4 &= M_4 - 3M_2^2 \\ I_5 &= M_5 - 10M_2M_3 & I_6 &= M_6 - 15M_2M_4 - 10M_3^2 + 30M_2^3 \end{aligned}$$

How to derive these? If you are a little bit sophisticated about probability (from a previous course?) look at Wikipedia "Cumulant". It's very cool!

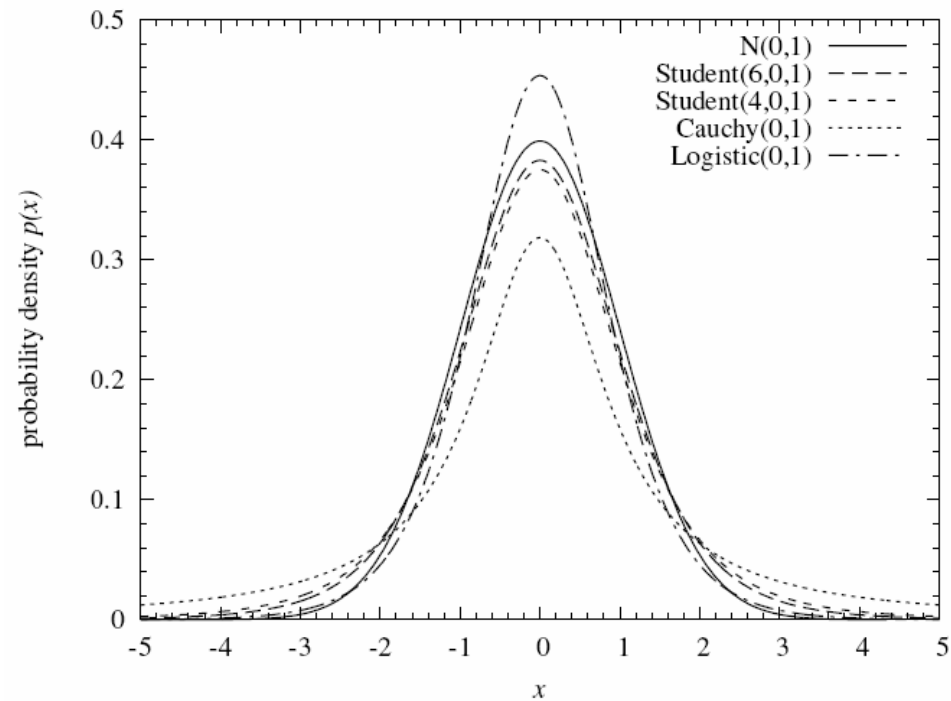
Skew and kurtosis are dimensionless combinations of semi-invariants

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

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.

Let us review some standard (i.e., frequently occurring) distributions:

The “bell shaped” ones differ qualitatively by their tail behaviors:





Normal (Gaussian) has the fastest falling tails:

$$x \sim N(\mu, \sigma), \quad \sigma > 0$$
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \left[\frac{x - \mu}{\sigma}\right]^2\right)$$

Cauchy (aka Lorentzian) has the slowest falling tails:

$$x \sim \text{Cauchy}(\mu, \sigma), \quad \sigma > 0$$
$$p(x) = \frac{1}{\pi\sigma} \left(1 + \left[\frac{x - \mu}{\sigma}\right]^2\right)^{-1}$$

Cauchy has area=1 (zero<sup>th</sup> moment), but no defined mean or variance (1<sup>st</sup> and 2<sup>nd</sup> moments divergent).

Student has power-law tails:

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

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

“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  $\sigma$  is not (quite) the standard deviation:

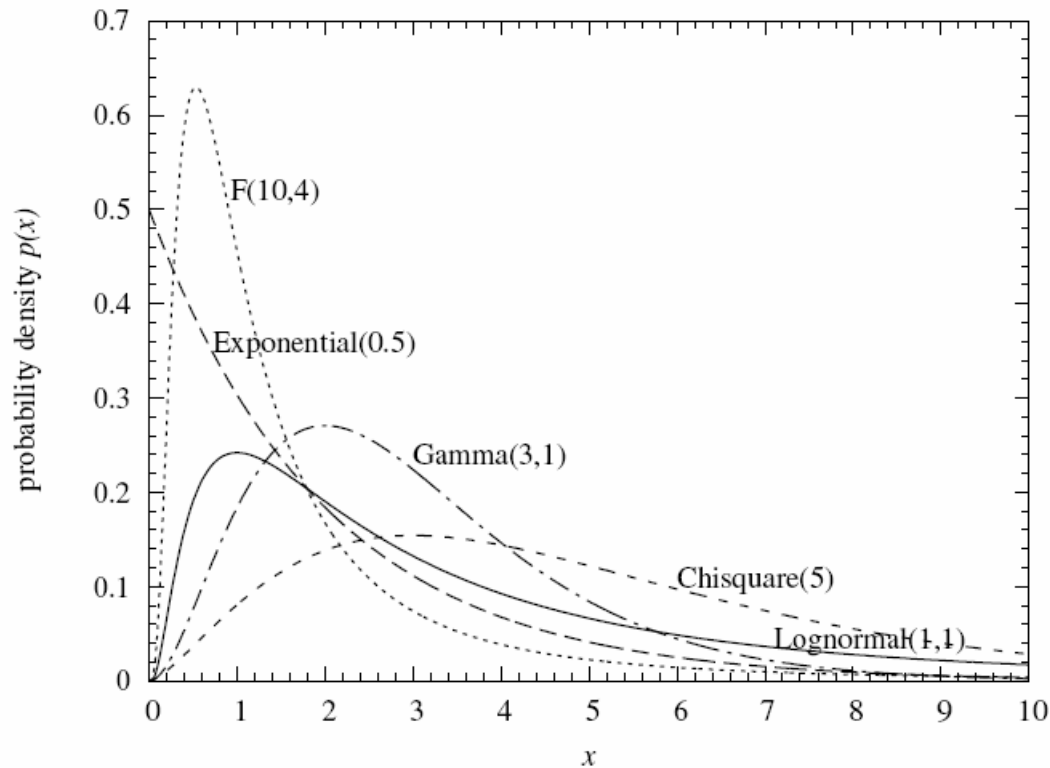
$$\text{Var}\{\text{Student}(\nu, \mu, \sigma)\} = \frac{\nu}{\nu - 2} \sigma^2$$

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. Brewing was one of the first “exact” modern manufacturing processes. More on Student later...



Another class of distributions model positive quantities:



Exponential:

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

## Lognormal:

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

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  $\mu$  and  $\sigma$  are the mean and standard deviation in  $\log x$  space, they are *not* so in  $x$  space. Rather,

$$\text{Mean}\{\text{Lognormal}(\mu, \sigma)\} = e^{\mu + \frac{1}{2}\sigma^2} \quad (6.14.32)$$
$$\text{Var}\{\text{Lognormal}(\mu, \sigma)\} = e^{2\mu} e^{\sigma^2} (e^{\sigma^2} - 1)$$

```
p = (1 / Sqrt[2 Pi]) (1 / (sig x)) Exp[-(1 / 2) * (Log[x] - mu) ^2 / sig ^2]
```

$$\frac{e^{-\frac{(-\mu + \log[x])^2}{2 \text{sig}^2}}}{\sqrt{2 \pi} \text{sig} x}$$

Mathematica (and also MATLAB) can do these integrals, no problem!

```
moments = Integrate[{1, x, x^2} p, {x, 0, Infinity}, Assumptions -> sig > 0,  
GenerateConditions -> False]
```

$$\left\{1, e^{\mu + \frac{\text{sig}^2}{2}}, e^{2(\mu + \text{sig}^2)}\right\}$$

```
Simplify[moments[[3]] - moments[[2]]^2]
```

$$e^{2\mu + \text{sig}^2} (-1 + e^{\text{sig}^2})$$

## Gamma distribution:

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

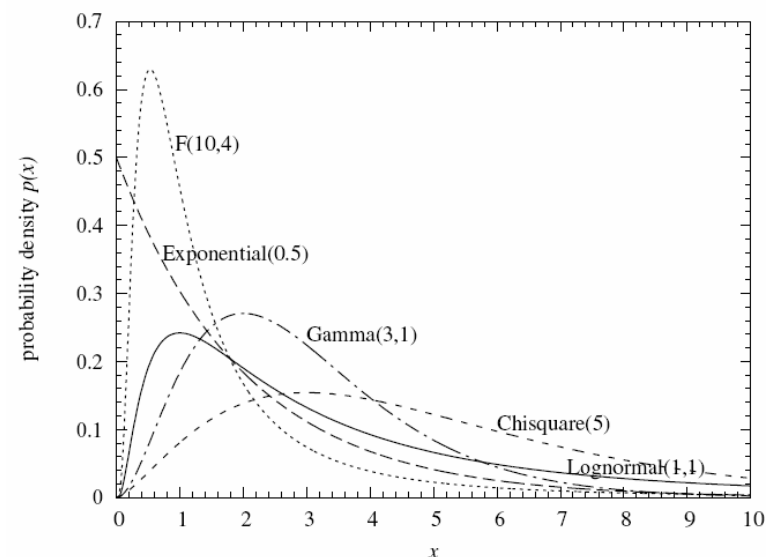
$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0$$

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

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

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  $\nu$  that determines both peak location and width.  
 $\nu$  is often an integer, called “number of degrees of freedom” or “DF”

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

the independent variable is  $\chi^2$ , not  $\chi$

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

It's actually just a special case of Gamma, namely  $\text{Gamma}(\nu/2, 1/2)$

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

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

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

Computationally, one wants efficient 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)

$$P(x) \equiv \int_{-\infty}^x p(x') dx'$$

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  $\mu$  and  $\sigma$ . 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)**