distributions

Ben Bolker

04 Sep 2018

Notation

I will try to follow the following notation:

- *scalars, estimates*: lower-case roman: a, b
- *vectors*: lowercase bold math Roman: $\boldsymbol{y}$
- *matrices*: bold math Roman: $\boldsymbol{X}$
- *random variables*: upper-case Roman: Y
- *estimates* (alternative): "hat": $\hat{\beta}$
- *model parameters* lower-case Greek: β (scalar), $\boldsymbol{\beta}$ (vector), β_i (vector element)

Probability distributions will be written out as proper, Roman names, possibly abbreviated: Normal, Beta, NegBinom, Gamma (the Gamma *function* is also spoken "Gamma", but is written as $\Gamma(x)$)

The symbol $\sim$ means "distributed as": $Y \sim \text{Normal}(\mu, \sigma^2)$

Distributions

Related to the Normal distribution

These are mostly used *not* used to describe data, but rather as theoretical constructs (e.g. null distributions, Bayesian priors). *Non-central* variants are mostly used for power analyses.

- Normal: standard (Z: $\mu = 0$, $\sigma = 1$), non-standard, MVN: closed under convolution (addition of random variables)
- χ^2: central, non-central: also closed. Central: mean n, var $2n$. Non-central; mean $n + \lambda$, var $2n + 4\lambda$, where $\lambda = \sum \mu_i^2$.
- MVN has $\boldsymbol{y}^T \boldsymbol{V}^{-1} \boldsymbol{y} \sim \chi_k^2$
- (Wishart distribution $W(\boldsymbol{V}, n)$: distribution of $\sum_{i=1}^N \boldsymbol{y}_i \boldsymbol{y}_i^T$ where the individual vectors are $\text{MVN}(0, \boldsymbol{V})$)
- (Student) t: $Z/\sqrt{X^2/n}$
- F: $(X_1^2/n_1)/(X_2^2/n_2)$ (central, non-central)

Matrix rules/quadratic forms:

Positive definiteness

- $\leftrightarrow$ positivity of quadratic form ($\boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y} > 0$ when $\boldsymbol{y}$ is not all zero)
- $\rightarrow$ all positive eigenvalues (variances)
- $\rightarrow$ invertible

Singular matrices: non-full-rank (quadratic forms have χ^2 distribution with lower df)

Others (exponential family etc.)

- Binomial: counts with known denominator (beta-binomial). Closed under convolution if p is homogeneous.
- Poisson: counts. $\exp(-\lambda)\lambda^x/(x!)$ (Can sometimes model proportions via Poisson with offset.) Closed under convolution. Variance $=$ mean. Limit of binomial as $N \to \infty$, $p \to 0$ with $\lambda = Np$.
- Negative binomial: can be described a discrete waiting time distribution ($\propto p^n(1-p)^x$, with mean $n(1-p)/p$) **or** as an overdispersed (Gamma-Poisson) count distribution $\propto (k/(k+\mu))^k(\mu/(k+\mu))^x$ (in R, must specify `mu=` explicitly)
- Gamma (exponential): waiting-time distributions. $\frac{1}{s^a\Gamma(a)}x^{a-1}\exp(-x/s)$. Mean = as, variance = as^2, coefficient of variance = $1/\sqrt{a}$ χ^2_n $=$ Gamma$(s=2, a=n/2)$. Note Gamma vs Γ.

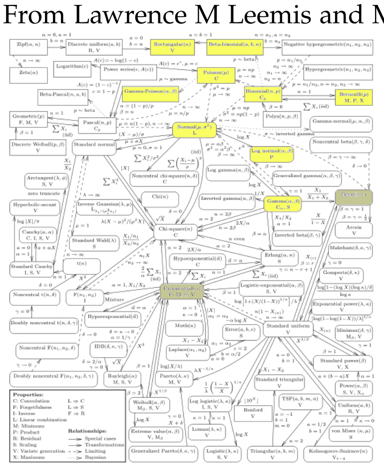
Exponential family:

$$\boxed{f(y;\theta,\phi) = \exp[(a(y)b(\theta)+c(\theta))/f(\phi)+d(y,\phi)]}$$

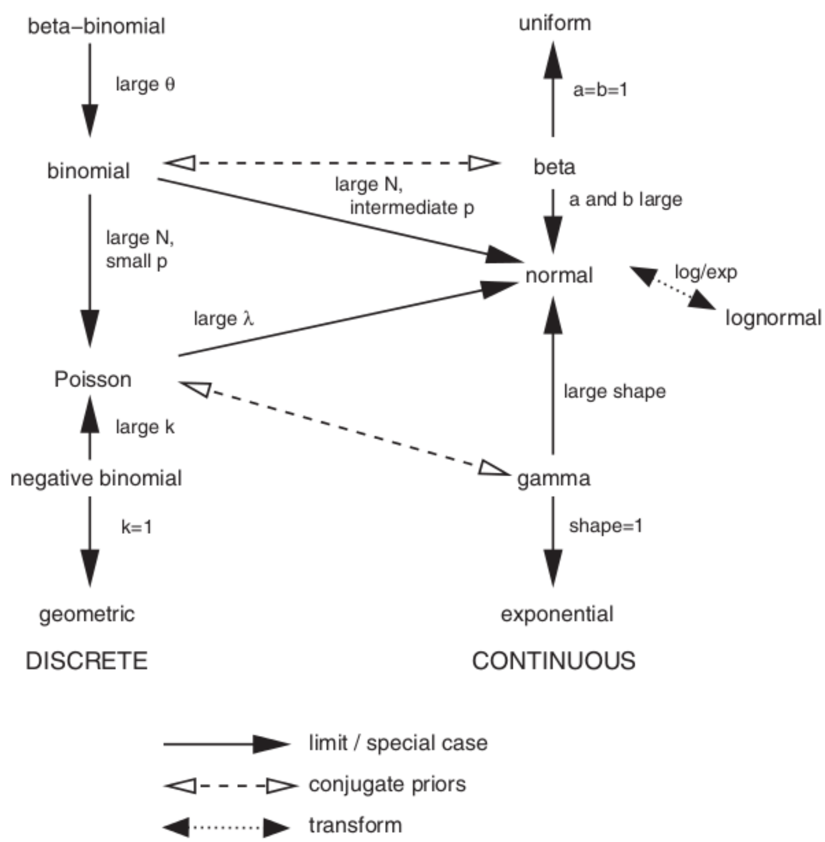
e.g. Poisson (with $\lambda \to \theta$, $x \to y$) ($\phi = 1$):

$$\begin{aligned} f(y,\theta) &= \exp(-\theta)\theta^y/(y!) \\ &= \exp\left(\underbrace{y}_{a(y)}\underbrace{\log\theta}_{b(\theta)} + \underbrace{(-\theta)}_{c(\theta)} + \underbrace{(-\log(y!))}_{d(y)}\right) \end{aligned}$$

From Lawrence M Leemis and McQueston (2008) :



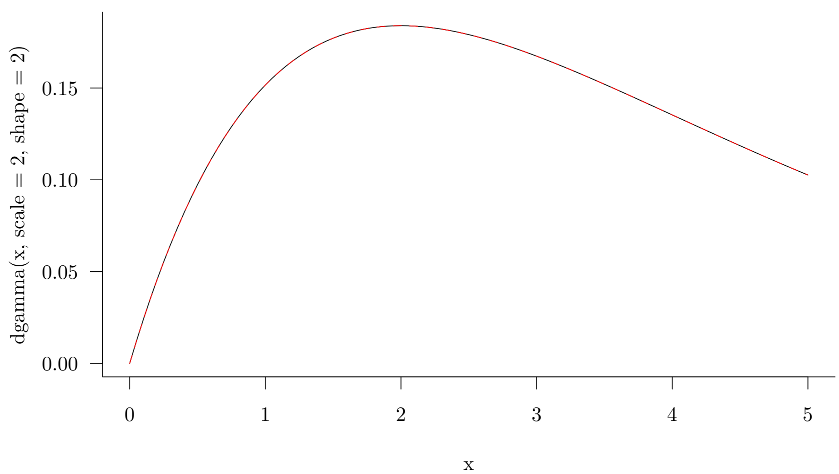
(also see interactive/corrected version: Lawrence M. Leemis et al. (2012))


beta−binomial
large θ
binomial
large N, small p
Poisson
large k
negative binomial
k=1
geometric
DISCRETE
uniform
a=b=1
beta
a and b large
large N, intermediate p
normal
log/exp
lognormal
large λ
large shape
gamma
shape=1
exponential
CONTINUOUS
limit / special case
conjugate priors
transform

Distributions in R

- `d*`, `p*`, `q*`, `r*` functions
  `binom`, `pois`, `nbinom`, `gamma`, `chisq`, **`mvrnorm`)
- base package: `?Distributions`
- Distributions task view; `SuppDists` package; `mvtnorm` package
- `distr`, `distrDoc` packages for general operations on distributions (convolutions etc.).
- "Lazy math: with R: e.g.

```
par(las=1,bty="l")
curve(dgamma(x,scale=2,shape=2),from=0,to=5)
curve(dchisq(x,df=4),add=TRUE,col=2,lty=2)
```



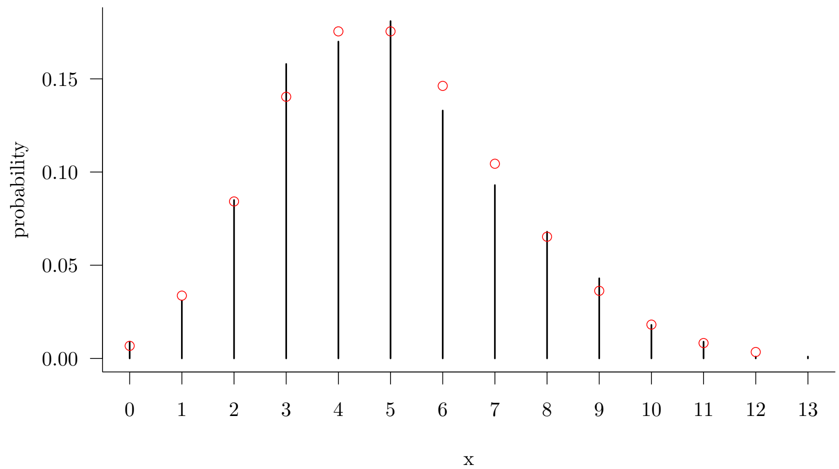
or

```
set.seed(101)
var(rnbinom(10000,mu=1,size=2))

## [1] 1.421275
```

or

```
par(las=1,bty="l")
x <- rpois(1000,lambda=2)+rpois(1000,lambda=3)
plot(prop.table(table(x)),ylab="probability")
## for continuous distributions: hist(x,freq=FALSE,breaks=100,col="gray")
curve(dpois(x,5),from=0,to=12,n=13,add=TRUE,type="p",col=2)
```



References

Leemis, Lawrence M, and Jacquelyn T McQueston. 2008. "Univariate Distribution Relationships." *The American Statistician* 62 (1): 45–53. doi:10.1198/000313008X270448.

Leemis, Lawrence M., Daniel J. Luckett, Austin G. Powell, and Peter E. Vermeer. 2012. "Univariate Probability Distributions." *Journal of Statistics Education* 20 (3): null. doi:10.1080/10691898.2012.11889648.