9/24-26: Expectation and Distributions

Outline for Today

- Brief comments on decision theory
- Special distributions
- A law of large numbers

Brief look at Decision Theory

Statistical decision theory – making decisions in the presence of statistical knowledge.

Example (Berger): A drug company is deciding whether or not to market a new pain reliever. Two important factors:

1. Proportion of people θ_1 for whom the drug will be effective
2. Market share θ_2 the drug will capture

$\theta = (\theta_1, \theta_2)$ are unknown, but the company needs to decide whether to market the drug, the price, etc.

Decision Theory

For each in a set of actions $a \in \mathcal{A}$, if the parameter is θ, a *loss* $L(a, \theta)$ is associated with choosing action a.

The *risk* is the expected loss:

$$R = \int_\Theta L(a, \theta)\, dF(\theta)$$

and one chooses the action that minimizes the risk.

Note: Can replace θ by x, the value of a random variable X.

$$R = \int_{-\infty}^{\infty} L(a, x)\, dF_X(x)$$

Decision Theory

Suppose that the company wants to estimate market share θ_2.

The "action" chosen is to use a certain estimate of this in further management decisions.

Suppose

$$L(\theta_2, a) = \begin{cases} \theta_2 - a & \text{if } \theta_2 - a \geq 0, \\ 2(a - \theta_2) & \text{if } \theta_2 - a \leq 0 \end{cases}$$

An under estimate is penalized more than an overestimate

Decision Theory

Suppose that the company does a study, interviews n people and finds X people would buy the drug.

Assume $X = \text{Binom}(n, \theta_2)$. Then

$$f(\theta_2 \,|\, x) \;\propto\; \binom{n}{x} \theta_2^x \,(1 - \theta_2)^{n-x} \, f(\theta_2)$$

$f(\theta_2)$ might be affected by previous drugs marketed, etc., and is very important in this case.

Utility and Decision Theory

Sometimes more convenient to use "utility" $U(\theta, a)$; can take loss $L(\theta, a) = -U(\theta, a)$.

Canonical example: X outcome of some gambling event, say a lottery—person gains X dollars. How much is their ticket worth to them?

Let $U(X)$ denote the "utility" of winning X dollars. The expected utility (negative risk) is $E[U(X)]$.

If ticket is sold for t dollars, gain is $U(t)$. Value of the ticket is that t for which

$$U(t) \;>\; E[U(X)]$$

Utility and Decision Theory

Suppose $U(x) = x^2$ and ticket wins \$36 with probability $\frac{1}{4}$ and wins \$0 with probability $\frac{3}{4}$.

Then the value of the ticket to the holder (having utility U) is

$$\begin{aligned} t^2 &= E[U(X)] \\ &= \frac{1}{4} \cdot 36^2 + \frac{3}{4} \cdot 0 \\ &= 324 \end{aligned}$$

so $t = 18$.

Utility and Decision Theory

If $U(x) = \sqrt{x}$ then the value of the ticket is

$$\begin{aligned} \sqrt{t} &= E[U(X)] \\ &= \frac{1}{4} \cdot \sqrt{36} + \frac{3}{4} \cdot 0 \\ &= 1.5 \end{aligned}$$

so $t = 2.25$.

(IMHO, such uses of decision theory are highly suspect, since the utility function can never be effectively set; same problem with models of rational behavior in economics...)

Example from Information Retrieval

1. Two parts of IR problem: modeling documents and queries
2. Making a decision on what documents to present to the user

Naturally cast in framework of statistical decision theory.

(L. and Zhai, SIGIR 2001) and (Zhai thesis, 2002).

Distributions

Let's now take a closer look at some of the standard parametric distributions.

Poisson

Probability mass function:

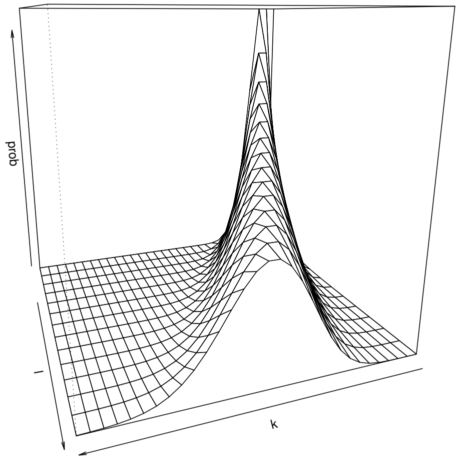
$$f(k \mid \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Mean:

$$\begin{aligned} \sum_k k\, f(k \mid \lambda) &= \sum_{k \geq 0} k \frac{e^{-\lambda}\lambda^k}{k!} \\ &= \lambda \sum_{k \geq 1} \frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} \\ &= \lambda \end{aligned}$$

Variance $= \lambda$ by a similar calculation.

Poisson



Poisson Process

If objects "arrive" in such a way that

1. number of arrivals in disjoint time intervals are independent

2. interarrival times are exponential with rate λ

then the number arrivals in time t is Poisson(λt).

Arrivals are said to form a *Poisson process*.

Binomial vs. Poisson

For large n and small p, Binom$(n, p) \approx$ Poisson(np).

For f the binomial density,

$$\begin{aligned} f(k \mid n, p) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n \cdot (n-1) \cdots (n-k+1)}{k!} p^k (1-p)^{n-k} \\ &= \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \end{aligned}$$

letting $\lambda = np$.

Binomial vs. Poisson

Letting $n \to \infty$ and $p \to 0$ with np held constant, we get

$$\frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n} \left(1 - \frac{\lambda}{n}\right)^{-k} \longrightarrow 1$$

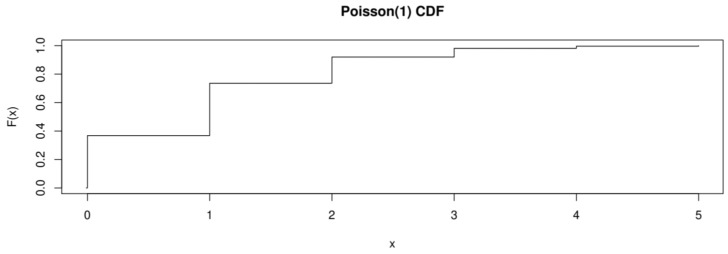
and

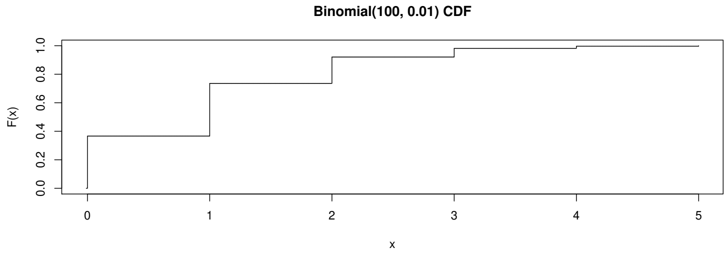
$$\left(1 - \frac{\lambda}{n}\right)^n \longrightarrow e^{-\lambda}$$

So,

$$f(k \mid n, p) \longrightarrow \frac{e^{-\lambda} \lambda^k}{k!}$$

Binomial vs. Poisson





Poisson and Multinomial

The Poisson can be used togeher with the multinomial in a way that can sometimes make calculation easier:

Suppose that we sample from a (high-dimensional) multinomial and then remove duplicates. The probability of a bit vector $\mathbf{x}$ is

$$\mathrm{Pr}(\mathbf{x}) \;=\; \sum_n \mathrm{Pr}(n) \sum_{n_1>0} \cdots \sum_{n_m>0} \binom{n}{n_1 \cdots n_m} \prod_{i=1}^m \theta(x_i)^{n_i}$$

By assuming that $n \sim \mathrm{Poisson}(\lambda)$, this simplies to

$$\mathrm{Pr}(\mathbf{x}) \;=\; e^{-\lambda} \prod_{i=1}^m \left(e^{-\lambda\,\theta(x_i)} - 1\right)$$

Geometric

Prob. mass function $f(k \mid p) = (1-p)^k p$. Using the differentiation trick for the geometric series, can compute

$$E(X) \;=\; \frac{1-p}{p}$$
$$\mathrm{Var}(X) \;=\; \frac{1-p}{p^2}$$

The geometric distribution has the *memoryless property*

$$\mathrm{Pr}(X = k+t \mid X \geq k) \;=\; \mathrm{Pr}(X = t)$$

Negative Binomial

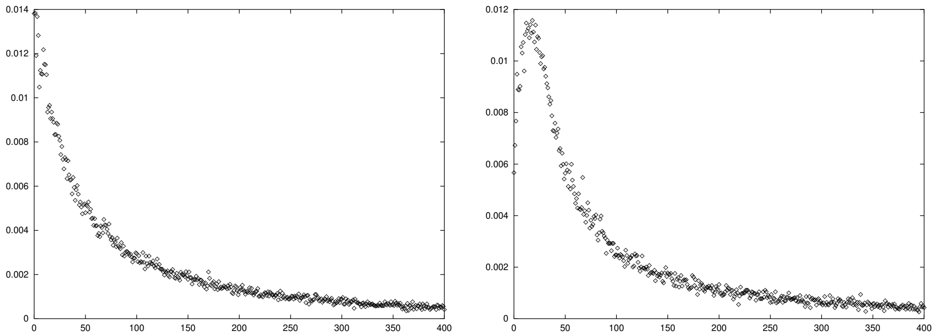
Recall the negative binomial with parameters r and p: outcome of experiment where we record the number of failures before the r-th success in sequence of iid Bernoulli trials.

$$\text{NegBinom(r,p)} \;=\; \underbrace{\mathrm{Geom}(p) \star \mathrm{Geom}(p) \star \cdots \star \mathrm{Geom}(p)}_{r \text{ times}}$$

So, if $X \sim$ NegBinom(r,p) then

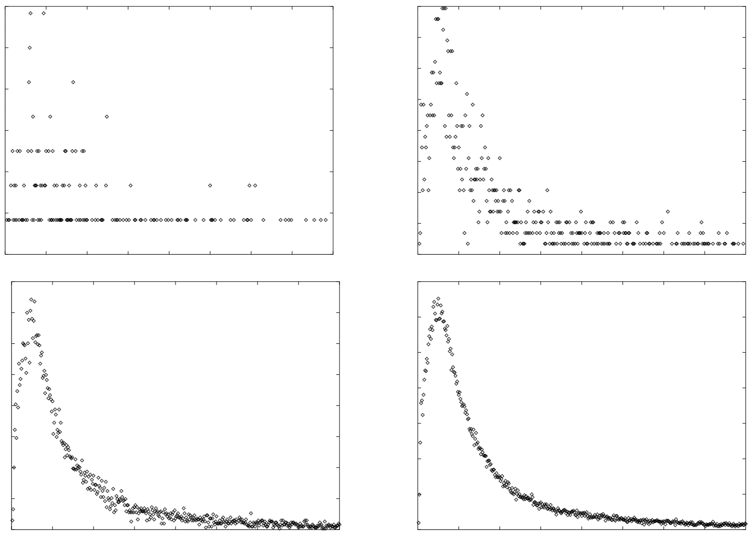
$$E(X) \;=\; r\,\frac{1-p}{p}$$
$$\mathrm{Var}(X) \;=\; r\,\frac{1-p}{p^2}$$

Example – Word Spacings



For a given distance k on the x-axis, the value on the y-axis is the empirical probability that two "trigger" words within the group are separated by exactly $k + 2$ words, conditional on the event that they co-occur within a 400 word window.

Example – Word Spacings



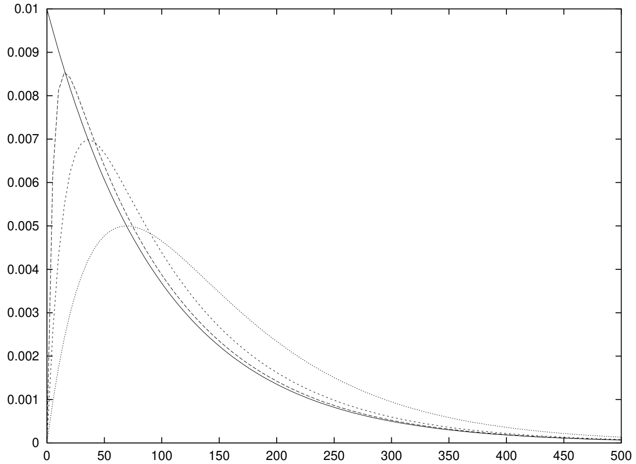
Example – Word Spacings

Three characteristics:

1. Mode away from zero
2. Exponential fall off
3. Long tail

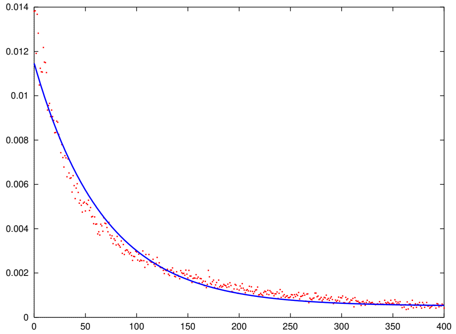
What type of distribution will fit this data?

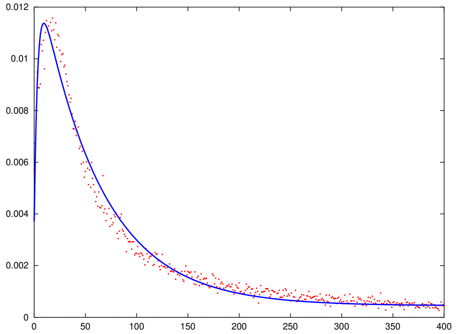
Example – Word Spacings



A sequence of two-stage exponential models $p_{\mu_1,\mu_2}(x)$ with $\mu_1 = 0.01,\ 0.02,\ 0.06,\ 0.2,\ \infty$ and $\mu_2 = 0.01$.

Example – Word Spacings





The empirical distance distributions fit to a three-parameter mixture model $p_{\mu_1,\mu_2,\alpha}$.

Normal Distribution

By far the most important in probability and statistics. Reasons:

1. Central limit theorem
2. Squared error – linear algebra operations often have interpretation in terms of Gaussian
3. Calculation
4. Often matches physics/empirical – example: channel models for communications, error correcting codes

Normal Distribution

Normal distribution is invariant under linear transformations.

If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$ then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

If X_i are independent distributed as $\mathcal{N}(\mu_i, \sigma_i^2)$, then $X_1 + X_2 + \cdots + X_k \sim N(\mu_1 + \cdots + \mu_k, \sigma_1^2 + \cdots + \sigma_k^2)$

Example

Suppose that $X \sim \mathcal{N}(5, 4)$. What is Pr$(1 < X < 8)$?

$$\begin{aligned}
\Pr(1 < X < 8) &= \Pr\left(\frac{1-5}{2} < \frac{X-5}{2} < \frac{8-5}{2}\right) \\
&= \Pr(-2 < Z < 1.5) \\
&= \Phi(1.5) - \Phi(-2) \\
&= \Phi(1.5) - (1 - \Phi(2)) \\
&\approx 0.9105
\end{aligned}$$

Example

Suppose that a sample of size n is taken from a $\mathcal{N}(\mu, 9)$ distribution. What is the minimum n for which

$$\Pr\left(\left|\overline{X}_n - \mu\right| \leq 1\right) \quad \geq \quad 0.95\,?$$

Since $Z = \frac{\sqrt{n}}{3}\left(\overline{X}_n - \mu\right)$,

$$\Pr\left(\left|\overline{X}_n - \mu\right| \leq 1\right) \quad = \quad \Pr\left(|Z| \leq \frac{\sqrt{n}}{3}\right) \geq 0.95$$

Calculating using Φ gives $\frac{\sqrt{n}}{3} \geq 1.96$ or $n \geq 35$.

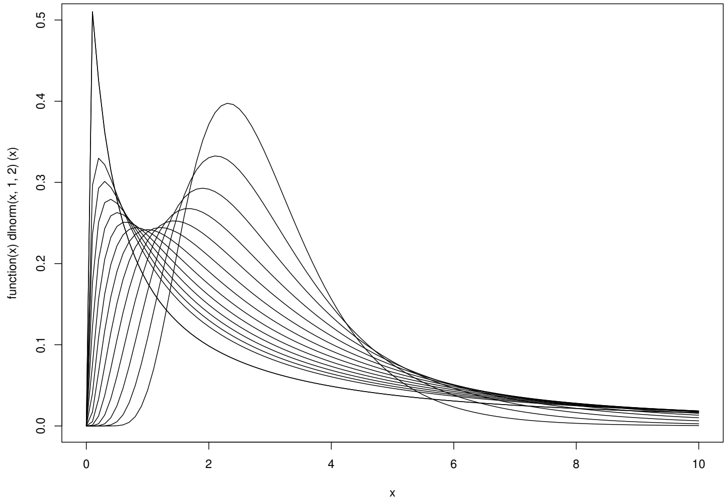
Lognormal

X has a *Lognormal* distribution with parameters μ, σ^2 in case $\log(X) \sim \mathcal{N}(\mu, \sigma^2)$.

So,

$$\begin{aligned}\Pr(X \leq x) \quad &= \quad \Pr\left(\log(X) \leq \log(x)\right) \\ &= \quad \Phi\left(\frac{\log(x) - \mu}{\sigma}\right)\end{aligned}$$

Lognormal



Gamma

The gamma is to the exponential as the negative binomial is to the geometric (convolution)

Gamma

$X \sim \text{Gamma}(\alpha, \beta)$ if the density is given by

$$f(x \mid \alpha, \beta) \;=\; \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-\beta x}$$

Mean: $\dfrac{\alpha}{\beta}$

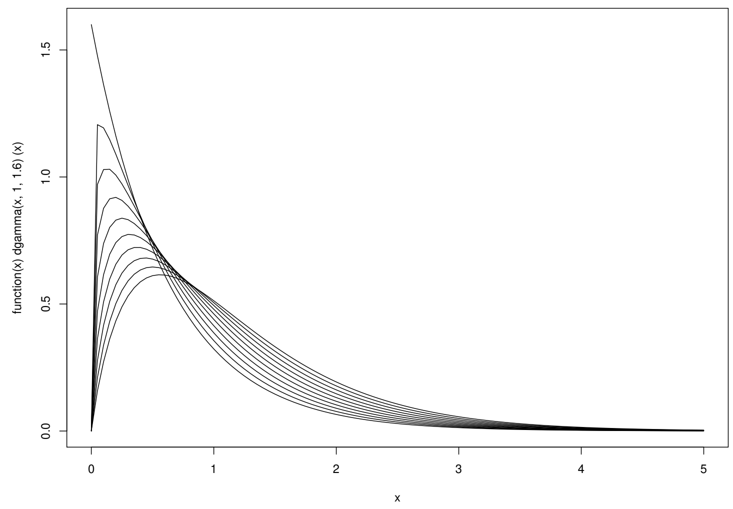
Variance: $\dfrac{\alpha}{\beta^2}$

Gamma

If $X_i \sim \text{Gamma}(\alpha_i, \beta)$ then $\sum X_i \sim \text{Gamma}(\sum \alpha_i, \beta)$.

Think of the case of a convolution of exponentials: $X_i \sim \text{Gamma}(1, \beta)$. We are sending a "customer" through a series of queues, each being emptied at a rate β.

Gamma



Beta

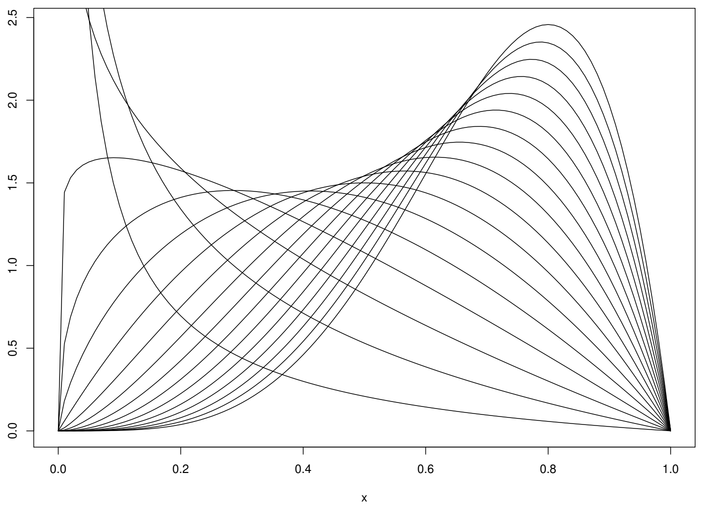
The Beta distribution is on the unit interval $[0, 1]$. Think of $f(x \mid \alpha, \beta)$ as the probability the coin we're flipping has bias x.

$$f(x \mid \alpha, \beta) \;=\; \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1}$$

Can think of α and β are "imaginary" numbers of times we've observed heads and tails.

Mean: $\dfrac{\alpha}{\alpha+\beta}$

Beta



Dirichlet

The Dirichlet is the analogue of the Beta for a k-sided coin.

$X \sim \mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_k)$ then

$$f(x \mid \alpha_1, \ldots, \alpha_k) \;=\; \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

A distribution on the $(k-1)$ simplex.

Mean: $E[x_i] = \dfrac{\alpha_i}{\sum_j \alpha_j}$