so that $\ln \Gamma(x) \doteq \frac{1}{2}\ln(2\pi) + (x-\frac{1}{2})\ln x - x.$

$X \sim \Gamma(\alpha, \beta)$ | $\psi_X(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha}$ for $t < \beta$

so $E(X) = \frac{\alpha}{\beta}$ and $V(X) = \frac{\alpha}{\beta^2}$ | $SD(X) = \frac{\sqrt{\alpha}}{\beta}$

Alternative expression | $\psi_X(t) = \left(\frac{\beta}{\beta - t}\right)^{\alpha}$ for $t < \beta$

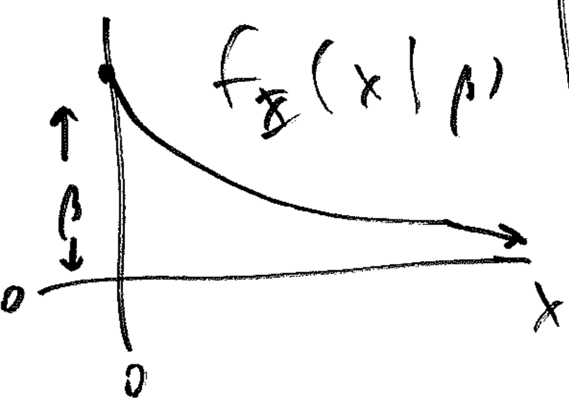
Special case of $\Gamma(\alpha, \beta)$ | With $\alpha = 1$ the PDF is

$$f_X(x \mid \beta) = \beta e^{-\beta x} I(x > 0)$$

But this is just our old friend the Exponential distribution.

$X \sim \text{Exponential}(\beta)$ $\qquad \psi_X(t) = \frac{\beta}{\beta - t}, \; t < \beta$

$E(X) = \frac{1}{\beta}$ $\qquad V(X) = \frac{1}{\beta^2}$ $\qquad SD(X) = \frac{1}{\beta}$



~~Notice that the Exponential distribution has S TMR equal to $\frac{1}{\beta}$; this suggests it's related somehow to the Poisson dist.~~

Theorem Suppose that arrivals (events) occur according to a Poisson process with rate β per unit time. Set Z_k = time until k^{th} arrival, $k = 1, 2, \ldots$ and define

$$Y_1 = Z_1 - 0$$

$$Y_2 = Z_2 - Z_1$$

$$\ldots \quad Y_k = Z_k - Z_{k-1} \quad \text{for } k = 2, 3, \ldots$$

The Y_i are called the inter-arrival times. Then it turns out that $Y_i \overset{IID}{\sim}$ Exponential(β)

The Exponential dist. is also related to the Geometric dist., in that they both have a memoryless property

Theorem

$X \sim$ Exponential(β) ; $t > 0$, $h > 0$

$$\rightarrow P(X \geq t+h \mid X \geq t) = P(X \geq h)$$

Example X = time from initial use until a manufactured product fails (e.g., light bulb)

$F_X(x) = P(X \leq x)$. $1 - F_X(x) = P(X > x)$

$= P($"system survives" at least to time $x)$

For this reason, $1 - F_X(x)$ is called the survival function $S_X(x) = 1 - F_X(x)$ in medicine and the reliability function $R_X(x) = 1 - F_X(x)$ in engineering.

Earlier we showed that for $X \sim$ Exponential (β) $\rightarrow$ $F_X(x) = 1 - e^{-\beta x}$ for $x > 0$

So $S_X(x) = R_X(x) = e^{-\beta x}$ for this dist.

The instantaneous failure rate or hazard rate function is defined to be $H_X(x) = \dfrac{f_X(x)}{S_X(x)} = \dfrac{f_X(x)}{R_X(x)}$

This gives $P($ failure in interval $(x, x+\epsilon)$ $|$ survival to time $x)$ for small ϵ

Notice that if $X \sim$ Exponential (β) 280

then $$H_X(x) = \frac{\beta e^{-\beta x}}{e^{-\beta x}} = \beta \quad \left(\underline{\text{Constant in}} \; x\right)$$

The Exponential is the only failure rate distribution with <u>constant hazard</u>. Returning

---

to the earlier result that $X \sim$ Exponential (β),

$$\rightarrow P(X \geq t+h \mid X \geq t) = P(X \geq h),$$

---

for all $t > 0$, $h > 0$

this says that if the product has survived to time t, the chance it will survive to time $(t+h)$ is the same as the original chance of surviving from time 0 to time h; i.e., "the system doesn't remember how long it's survived" (this often makes the Exponential unrealistic in practice)

Consequences | ① $X_i \overset{IID}{\sim}$ Exponential(β) $(i=1,\dots,n)$,

then

$$Y_1 = \min(X_1, \dots, X_n) \sim \text{Exponential}(n\beta).$$

Beta Distribution | $\alpha, \beta > 0$ $X \sim \text{Beta}(\alpha, \beta) \leftrightarrow$

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \cdot \underbrace{I(0 < x < 1)}_{\text{support of } X}$$

The name comes from the normalizing constant: the function $x^{\alpha-1}(1-x)^{\beta-1}$ has no closed-form anti-derivative, so people just made a

Definition | For all $\alpha > 0$, $\beta > 0$

$$\underbrace{B(\alpha, \beta)}_{\text{beta function}} \doteq \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$$

Can show that $B(\alpha, \beta) = \dfrac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

(α, β) jointly control the shape of the Beta(α, β) dist.

(yuck)

$X \sim \text{Beta}(\alpha, \beta)$ | $\psi_X(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{t^k}{k!}$

$E(X) = \dfrac{\alpha}{\alpha+\beta}$ | $V(X) = \left(\dfrac{\alpha}{\alpha+\beta}\right)\left(\dfrac{\beta}{\alpha+\beta}\right)\left(\dfrac{1}{\alpha+\beta+1}\right)$

Case study (Castaneda v. Partida continued)

$n = 220$ grand jurors chosen from ( ) eligible population of Hidalgo County, Texas, which was 79.1% Mexican-American, but only $s = 100$ selected grand jurors were Mexican-American; let's summarize the information in a Bayesian fashion about evidence of discrimination.

Data S = # Mexican-Americans chosen in grand jury selection of $n = 220$ people

Unknown θ = actual probability of an eligible Mexican-American person being chosen $(0 < \theta < 1)$

Sampling model $(S \mid \theta) \sim \text{Binomial}(n, \theta)$,

i.e., $f_{S \mid \theta}(s \mid \theta)$ = $P(S = s \mid \theta) = \binom{n}{s} \theta^s (1 - \theta)^{n-s} \cdot I(s = 0, 1, \dots, n)$

PMF

Bayesian approach ① Information internal to data set about θ summarized by the likelihood (un-normalized) density, defined to be $\boxed{\ell(\theta \mid s) = c\, P(S = s \mid \theta)}$,

c an arbitrary positive constant — just think of $P(S = s \mid \theta)$ as a function of θ for fixed s.

Here $\ell(\theta \mid s) = c \binom{n}{s} \theta^s (1-\theta)^{n-s}$ can be absorbed into c since does not depend on θ

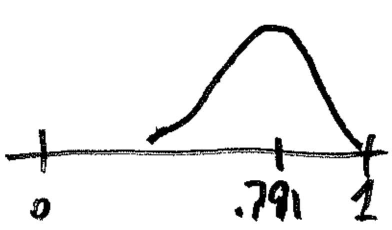
$= c\, \theta^s (1-\theta)^{n-s}$

(2) Information external to dataset about θ summarized by the prior density $f_{\Theta}(\theta)$. Here are some

possibilities for the prior, depending on your knowledge (information) base: (a) neutral prior $\theta \sim \text{Uniform}(0,1)$

this dist. embodies the information {θ could be anywhere between 0 and 1, with no value favored}

(b) cut the district attorney some slack prior



this prior gives the DA the benefit of the doubt when you're uncertain about what prior to use, write down all the reasonable priors & do a sensitivity analysis (use each prior one by one & see if posterior answer is essentially the same)

---

③ Combine internal & external information

with Bayes' Theorem

$$f_{\Theta|S}(\theta|s) = c \cdot f_{\Theta}(\theta) \cdot \ell(\theta|s)$$

$$\begin{pmatrix}\text{posterior}\\ \text{information}\end{pmatrix} = \begin{pmatrix}\text{normalizing}\\ \text{constant}\end{pmatrix} \cdot \begin{pmatrix}\text{prior}\\ \text{information}\end{pmatrix} \cdot \begin{pmatrix}\text{likelihood}\\ \text{information}\end{pmatrix}$$

Here

$$f_{\Theta|S}(\theta|s) = c\, f_{\Theta}(\theta)\, \theta^{s} (1-\theta)^{n-s}$$

Rev. Bayes himself noticed back in 1760

that if you take $f_{\Theta}(\theta) = c\,\theta^{\text{power}}(1-\theta)^{\text{power}}$

then the product of 2 such densities is another such density, meaning that the posterior would have the same form as the prior & likelihood, making calculations easier. Moreover, we already know the name of densities that look like $\theta^{\text{power}}(1-\theta)^{\text{power}}$:

the Beta distributions: $X \sim \text{Beta}(\alpha, \beta)$ $(\alpha > 0, \beta > 0) \rightarrow$

$$f_X(x) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

So let's take as our prior PDF $f_{\Theta}(\theta) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$

in the lawsuit case study; then

$$f_{\Theta|S}(\theta|s) = c\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right]\left[\theta^{s}(1-\theta)^{n-s}\right]$$

$$= c \, \theta^{(\alpha+s)-1} (1-\theta)^{(\beta+n-s)-1} = \text{Beta}(\alpha+s, \beta+n-s)$$

So the prior-to-posterior updating looks like this:

"Beta dist. is conjugate to the Binomial likelihood"

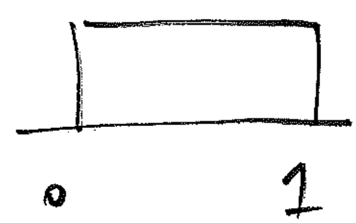
$$\left.\begin{array}{l} \theta \sim \text{Beta}(\alpha, \beta) \\ (S \mid \theta) \sim \text{Binomial}(n, \theta) \end{array}\right\} \rightarrow (\theta \mid s) \sim \text{Beta}(\alpha+s, \beta+n-s)$$

$s = 100$

$n = 220$

How choose (α, β)?

(a) Neutral prior



but $\text{Uniform}(0,1) = \theta^{1-1}(1-\theta)^{1-1}$

so $\theta \sim \text{Uniform}(0,1) \leftrightarrow \theta \sim \text{Beta}(1,1)$

(b) cut DA slack prior

There's an extremely useful thing that happens with conjugate priors:

$$\begin{array}{cl} \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \hline 1 \\ \vdots \\ 1 \end{array}\right] & \begin{array}{l} \beta \\ \\ \\ \alpha \end{array} \end{array}$$

pseudo data

prior effective sample size $(\alpha+\beta)$

mean $\frac{\alpha}{\alpha+\beta}$

$$\begin{array}{ccl} & y & \\ n & \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \hline 1 \\ \vdots \\ 1 \end{array}\right] & \begin{array}{l} n-s \\ \\ \\ s \end{array} \end{array}$$

sample data

data set sample size n

mean $\bar{y} = \frac{s}{n}$

Beta prior distribution acts like a dataset with α 1s & β 0s

with the property that if you do a Bayesian analysis with the Beta(α, β) prior and I do a frequentist analysis on the dataset with $(\alpha+s)$ 1s and $(\beta+n-s)$ 0s formed by merging the prior & sample data sets, we'll get the same results.

(b) cut the DA slack prior

mean of $\text{Beta}(\alpha, \beta)$ dist. is $\frac{\alpha}{\alpha+\beta}$; set this equal to 0.791

Suppose I want to put in prior information equivalent to a prior sample size $\frac{1}{10}$ as big as the data sample size (so $\frac{1}{10}$); set

$(\alpha+\beta) = \frac{1}{10} n = 22$; solve: set $\begin{cases} \alpha = 17.4 \\ \beta = 4.6 \end{cases}$

$n = 220$

$s = 180$

likelihood is

$$c\, \theta^s (1-\theta)^{n-s} = c\, \theta^{(s+1)-1} (1-\theta)^{(n-s+1)-1}$$

(a) Neutral prior: $\text{Beta}(1,1)$

prior sample size 2

$$= \text{Beta}(\underset{181}{s+1}, \underset{41}{n-s+1}) \text{ dist}$$

posterior is $\text{Beta}(\underset{181}{\alpha+s}, \underset{41}{\beta+n-s})$

(same as likelihood)

(b) Cut DA slack prior

Beta(17.4, 4.6) prior
α β

posterior is Beta($\alpha + s$, $\beta + n - s$)

$s = 220$, $n = 100$

117.4 124.6

prior	posterior mean	posterior SD
neutral	0.455	0.0333
Cut DA slack	0.485	0.0321

Posterior mean of θ is

$$\frac{\alpha + s}{\alpha + \beta + n}$$

Posterior SD is

$$\sqrt{\left(\frac{\alpha + s}{\alpha + \beta + n}\right)\left(\frac{\beta + n - s}{\alpha + \beta + n}\right)\left(\frac{1}{\alpha + \beta + n + 1}\right)}$$

The no-discrimination rate of 0.791 is

$$\frac{0.791 - 0.455}{0.0333} = 10.1$$

posterior SDs away from posterior expectation under the neutral prior and

$$\frac{0.791 - 0.485}{0.0321} \doteq 9.5 \text{ posterior SDs}$$

away from posterior expectation under the cut-DA-slack prior; there was discrimination Q.E.D.

Multinomial Distributions (back to discrete)

You're contemplating a population that contains elements of $k \geq 2$ types (e.g., {Democrat, Republican, Libertarian, Independent, Green, other}). Suppose the proportion of elements of type i is $0 < p_i < 1$ with $\sum_{i=1}^{k} p_i = 1$; $\boldsymbol{p} = (p_1, \dots, p_k)$.

You take an IID sample of size n from this pop.; X_i = # elements of type i in your sample; $\sum_{i=1}^{k} X_i = n$.

Can show that the vector $\underset{\sim}{X} = (X_1, \dots, X_k)$

has PMF

$$f_{\underset{\sim}{X} | n, \underset{\sim}{p}}(\underset{\sim}{x} | n, \underset{\sim}{p}) = \begin{cases} \binom{n}{x_1, \dots, x_k} p_1^{x_1} \cdots p_k^{x_k} & \text{if } \sum_{i=1}^{k} x_i = n \\ 0 & \text{else} \end{cases}$$

$$\left(\sum_{i=1}^{k} p_i = 1 \right)$$

where

$$\binom{n}{x_1, \dots, x_k} \triangleq \frac{n!}{x_1! \, x_2! \cdots x_k!}$$ is the multinomial coefficient

---

This is called the Multinomial $(n, \underset{\sim}{p})$ distribution.

$E(X_i) = n p_i$ | $V(X_i) = n p_i (1 - p_i)$

(just like binomial)

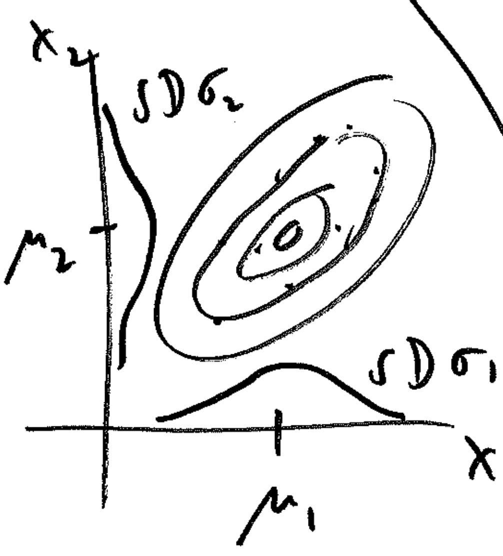
But now something new:

$C(X_i, X_j) = -n p_i p_j$

negatively correlated because $\sum_{i=1}^{k} X_i = n$.

Bivariate Normal Dist.

Can build a 2-dimensional (bivariate) version of the Normal dist. as follows:



$Z_1, Z_2 \overset{IID}{\sim} N(0, 1)$

Specify 5 parameters:

$-\infty < \mu_1 < +\infty$	$0 < \sigma_1 < \infty$
$-\infty < \mu_2 < +\infty$	$0 < \sigma_2 < \infty$

$-1 < \rho < +1$

Now build (X_1, X_2) with the transformation
$$X_1 = \mu_1 + \sigma_1 Z_1$$

$$X_2 = \sigma_2 \left[\rho Z_1 + \sqrt{1-\rho^2}\, Z_2 \right] + \mu_2$$

The joint PDF of $\underset{\sim}{X} = (X_1, X_2)$ is

then
$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}\,\sigma_1\sigma_2} \cdot \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right] \right\}$$

standard units

This is the Bivariate Normal $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ dist.

Easy to show that $E(X_1) = \mu_1$,

$E(X_2) = \mu_2$, $V(X_1) = \sigma_1^2$, $V(X_2) = \sigma_2^2$,

$\rho(X_1, X_2) = \rho$.

$\boxed{\text{Consequences of this Def.}}$

① $(X_1, X_2) \sim$ Bivariate Normal $\longrightarrow$

$$\begin{pmatrix} X_1, X_2 \\ \text{independent} \end{pmatrix} \longleftrightarrow \begin{pmatrix} X_1, X_2 \\ \text{uncorrelated} \end{pmatrix}$$

we already knew the $\rightarrow$ direction in general; what's new here is that correlation 0 implies independence if $(X_1, X_2) \sim$ Bivariate Normal.

② $(X_1, X_2) \sim$ Bivariate Normal$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$

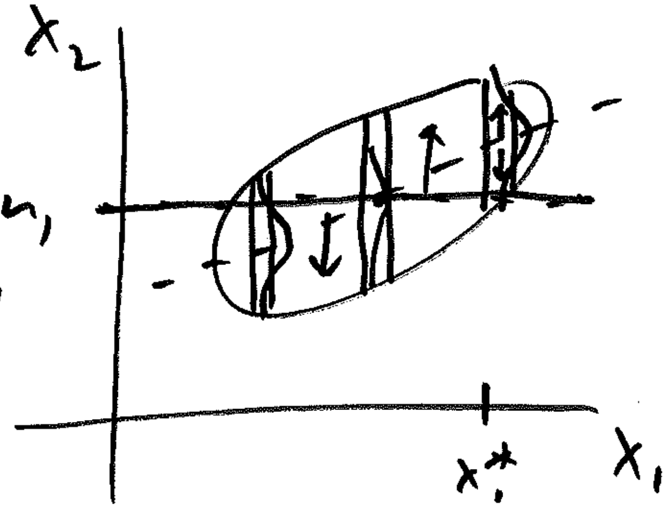
→ conditional distribution of X_2 given that $X_1 = x_1$ is (univariate) normal with mean $E(X_2 | x_1) =$

$$\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$$

and variance $V(X_2 | x_1)$

$$= (1 - \rho^2)\sigma_2^2$$

Galton, revisited



Result ② above says that if (X_1, X_2) are Bivariate Normal then the conditional distributions of X_2 given $X = x_1^*$ in all of the vertical strips are also normal

And the means of all these normal distributions in the vertical strips are connected together by Galton's

<u>regression line</u> $\boxed{\hat{X}_2 = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(X_1 - \mu_1)}$.

This line has slope $\beta_1 = \rho \frac{\sigma_2}{\sigma_1}$ and "y"-intercept

$\beta_0 = \mu_2 - \beta_1 \mu_1$ $\boxed{\hat{X}_2 = \beta_0 + \beta_1 X_1}$

Moreover, we can now quantify an earlier insight:

$\boxed{\text{ignore } X_1}$, predict $(\hat{X}_2)_{\text{no } x_1} = \mu_2 = E(X_2)$

$\left(\begin{array}{c}\text{root mean} \\ \text{squared} \\ \text{error}\end{array}\right)$ (RMSE) of this prediction is

$$\sqrt{V(X_2)} = \sigma_2$$

use x_1 to predict x_2

predict $(\hat{x}_2)_{\text{use } x_1} = E(X_2 \mid X_1 = x_1)$

$= \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$

RMSE of this prediction is $\sqrt{V(X_2 \mid x_1)} = \sigma_2 \sqrt{1-\rho^2}$

Since $-1 < \rho < 1$, $\sigma_2 \sqrt{1-\rho^2} \leq \sigma_2$

with equality only when $\rho = 0$.

---

(3) $(X_1, X_2) \sim \text{BivariateNormal}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$,

$Y = a_1 X_1 + a_2 X_2 + b$, (a_1, a_2, b) arbitrary constants

$\rightarrow Y \sim N(a_1\mu_1 + a_2\mu_2 + b,\ a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2\rho\sigma_1\sigma_2)$.

Large Random Samples (DS ch. 6)

You draw an IID random sample $X_1, \dots, X_n$ from a population, with the goal of estimating the population mean $\mu = E(X_i)$.

We've already seen that, from a root mean squared error point of view, the sample mean $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ is the best you can do (in the absence of prior information). It would be nice if $\bar{X}_n$ approached the right answer μ as n increases; how to quantify that idea?

Two inequalities that help

Markov inequality

Suppose X is a non-negative r.v., i.e. $P(X \geq 0) = 1$. Then for all real $t > 0$,

$$P(X \geq t) \leq \frac{E(X)}{t} \quad (*)$$

(Attributed to Andrey Markov (1856-1922), a Russian mathematician who did pioneering work on stochastic processes)

(*) says that, if $E(X)$ is fixed, you can't move more & more probability out into the right tail beyond a certain point.

25 Aug 17

Draper
↑
Lehmann
↑
Neyman
↑
Sierpiński
↑
Voronoy
↑
Markov

Laplace
↓
Bartels → Lobachevsky → Brashman → Chebyshev