where \( F(x) = \frac{1}{2} \ln(2\pi) + (x - \frac{1}{2}) \ln x \). 

If \( X \sim \Gamma(\alpha, \beta) \) then \( Y = \frac{X}{\beta} \) has the density

\[
f_Y(t) = \frac{1}{\beta} (1 - \frac{t}{\beta})^{-\alpha} \quad \text{for} \quad t < \beta.
\]

So \( E(X) = \frac{\alpha}{\beta} \) and \( \text{Var}(X) = \frac{\alpha}{\beta^2} \).

Alternative expression: \( Y \sim \Gamma(\alpha, \beta) \) with \( Y = \frac{X}{\beta} \).

Special case of \( \Gamma(\alpha, \beta) \):

\[
f_X(x | \beta) = \beta x^{\alpha-1} e^{-\beta x} I(x > 0)
\]

But this is just our old friend the exponential distribution.
$X \sim \text{Exponential}\left(\beta\right)$

$\mathbb{E}(X) = \frac{1}{\beta}$

$\text{Var}(X) = \frac{1}{\beta^2}$

$\text{PDF}(X) = \beta e^{-\beta x}$

Notice that the Exponential distribution has $\text{SNR}$ equal to 1, i.e., it is the solution to some how to the Poison dist.

Theorem: Suppose that arrivals (events) occur according to a Poison process with rate $\beta$ per unit time. Set $\frac{\lambda}{2} k = \text{time until k\textsuperscript{th} arrival}$ and define $T_1 = \xi_1 - 0$

$T_2 = \xi_2 - \xi_1$

$
\vdots
$

$T_k = \xi_k - \xi_{k-1}$ for $k = 2, 3, \ldots$
The $X_i$ are called the inter-arrival times.

Then it turns out that $X_i \sim \text{Exponential}(\beta)$.

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

Theorem

$X \sim \text{Exponential}(\beta); \ t > 0, \ h > 0$

$P(\bar{X} \geq t+h \mid \bar{X} \geq t) = P(\bar{X} \geq h)$

Example

$X =$ time until a manufactured product fails (e.g., light bulb)

$F_X(x) = P(\bar{X} \leq x) = 1 - F_X(x) = P(\bar{X} > x)$

$= P(\text{"system surviving" at least to time } x)$
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 \( F_X(x) = 1 - e^{-\beta x} \) for \( X \sim \text{Exponential} (\beta) \) for \( x > 0 \).

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

The instantaneous failure or hazard rate function is defined to be \( h_X(x) = \frac{f_X(x)}{S_X(x)} = \frac{f_X(x)}{1 - F_X(x)} \).

This gives \( \frac{\text{probability of failure in interval}}{\text{survival}} \) for small \( \Delta \).
Notice that if \( X \sim \text{Exponential} (\beta) \),

then \( H_X (x) = \frac{\beta e^{-\beta x}}{e^{-\beta x}} = \beta \left( \frac{\text{constant}}{x} \right) \)

The Exponential is the only failure rate distribution with constant hazard.

Returning to the earlier result that \( X \sim \text{Exponential} (\beta) \),

\[ P(X \geq t + h \mid X \geq t) = P(X \geq h), \quad \text{for all} \quad t > 0, \quad h > 0 \]

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.
Consequences 1. $X_i \sim \text{Exponential} (\beta)$
\[ i = 1, \ldots, n \]

Then
\[ Y = \min (X_1, \ldots, X_n) \sim \text{Exponential} (n \beta). \]

**Beta Distribution**

$x, \beta > 0$  $X \sim \text{Beta} (x, \beta) \leftrightarrow$

\[ f(x) = \frac{\Gamma(x+\beta)}{\Gamma(x) \Gamma(\beta)} x^{x-1} (1-x)^{\beta-1} I(0 < x < 1) \]

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

**Definition** For all $x > 0$  $B(x, \beta) = \int_0^1 x^{x-1} (1-x)^{\beta-1} dx$
Can show that \( B(\alpha, \beta) = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)} \) (282)

\( \alpha, \beta \) jointly control the shape of the Beta \((\alpha, \beta)\) dist. (yuck)

\[ \mathbb{E}(X) = 1 + \sum_{k=0}^{\alpha-1} \frac{1}{k!} \frac{\alpha + \beta + k}{\alpha + \beta + k} \cdot \frac{\alpha}{\alpha + \beta} \cdot \frac{1}{k!} \]

\[ \text{Var}(X) = \left( \frac{\alpha}{\alpha + \beta} \right) \left( \frac{\beta}{\alpha + \beta} \right) \left( \frac{1}{\alpha + \beta + 1} \right) \]

Case Study:

\( n = 220 \) jurors chosen from \( \text{eligible population of Hidalgo County, Texas, which was 79.1\% Mexican-American} \)

(continued) selected from jurors were Mexican-American. It's let's summarize the information in a Bayesian fashion about evidence of discrimination.
Data: \( S = \# \text{ Mexican-American in jury selection of } n = 220 \text{ people chosen} \)

unknown \( \theta = \text{ actual probability of an eligible Mexican-American person being chosen} \) \( (0 < \theta < 1) \)

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

PMF: \( f(s \mid \theta) = \text{Binomial}(s \mid n, \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s} \)

Bayesian Information integral approach to data set about \( \theta \) summarized by the likelihood (un-normalized) density, defined to be \( p(\theta \mid s) = c \cdot P(s \mid \theta) \)

can arbitrary positive constant - think of \( P(s \mid \theta) \) as a function of \( \theta \) for fixed \( s \).
Here \( x(\theta|s) = c \begin{pmatrix} \binom{n}{s} \theta^s (1-\theta)^{n-s} \end{pmatrix} \) can be absorbed into \( c \) since

do not depend on \( \theta \).

to dataset about \( \theta \) summarized by the prior density \( f(\theta) \). Here are some possibilities for the prior, depending on your knowledge base:

- (i) neutral \( \propto \text{uniform}(\theta) \)
- (ii) 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 and do a sensitivity analysis (use each prior one by one & see if answer is the same)

3) Combine internal & external information with Bayes' Theorem

Here

\[ f(\theta | s) = c \cdot f(\theta) \cdot \mathcal{L}(\theta | s) \]

\[ \mathcal{L}(\theta | s) \]

Posterior Information = (normalizing constant) \cdot (prior Information) \cdot (likelihood Information)

Rev. Bayes himself realized back in 1760
that if you take \( f(\theta) = c \theta^{\alpha - 1} (1 - \theta)^{\beta - 1} \) the product of \( k \) 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 \( \theta^{\alpha - 1} (1 - \theta)^{\beta - 1} \):

\( \xi \sim \text{Beta}(\alpha, \beta) \) \( (\alpha > 0, \beta > 0) \) \( \Rightarrow \) Beta distributions $f(\xi) = c \xi^{\alpha - 1} (1 - \xi)^{\beta - 1}$ as our prior PDF.

So, let's take \( f(\theta) = c \theta^{\alpha - 1} (1 - \theta)^{\beta - 1} \) in the lawsuit case study; then

\[
\phi(\theta | x) = c \left[ \theta^{\alpha - 1} (1 - \theta)^{\beta - 1} \right] \left[ \theta^x (1 - \theta)^{n-x} \right]
\]
\[ c \theta (\alpha + s)^{-1} \left(1 - \theta \right) (\beta + n - s)^{-1} = \text{Beta} (\alpha + s, \\ \beta + n - s) \]

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

\[ \theta \sim \text{Beta} (\alpha, \beta) \]

\[ \gamma \sim \text{Beta} (\alpha + s, \\ \beta + n - s) \]

\[ S | \theta \sim \text{Binomial} (n, \theta) \]

---

\[ S = 100 \]

\[ n = 220 \]

**How choose \((\alpha, \beta)\)?**

\((\alpha = 1, \beta = 1)\) is a **neural prior**

but \([0, 1] = \theta^{\alpha-1} (1 - \theta)^{\beta-1} \]

\[ S \sim \text{Uniform} (0, 1) \]

So \(\theta \sim \text{Uniform} (0, 1) \leftrightarrow \theta \sim \text{Beta} (1, 1) \]

---

\((b) \) cut slack prior

There's an extremely useful

thing that happens with conjugate priors:
Beta prior distribution acts like a dataset with $d + 1$ and $\beta + 1$ 

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

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

if you do a Bayesian analysis with the Beta $(d, \beta)$ prior and I do a Frequentist analysis on the dataset with $(d + s)$ and $(\beta + n - s)$ or formed by merging the prior & sample data sets, we'll get the same results.
Suppose I want to put in information as the prior sample size \( \frac{a}{n} \).  Set this equal to 0.59.  Then

\[
\frac{1}{n} \cdot \frac{a}{n} \cdot \frac{1}{n} = 0.59
\]

Solve for \( n \) as

\[
x = \frac{12.4}{0.59}
\]

\( n = 21.1 \) as the prior sample size.

The Jeffreys prior is 0.75

\[
\frac{(a+\beta)}{\alpha + \beta} \text{ Beta}(1,1)
\]

\[
\begin{align*}
\beta &= 1.22 \\
\alpha &= 1.22 \\
\end{align*}
\]

Suppose \( x \), the actual sample size, is

\[
x = 17.4
\]

\( \beta = 2.16 \)

\( \alpha = 2.16 \)

The Jeffreys prior is

\[
\frac{(a+\beta)}{\alpha + \beta} \text{ Beta}(1,1)
\]

\[
\begin{align*}
\beta &= 1.22 \\
\alpha &= 1.22 \\
\end{align*}
\]

Suppose \( x \), the actual sample size, is

\[
x = 17.4
\]

\( \beta = 2.16 \)

\( \alpha = 2.16 \)

The Jeffreys prior is

\[
\frac{(a+\beta)}{\alpha + \beta} \text{ Beta}(1,1)
\]

\[
\begin{align*}
\beta &= 1.22 \\
\alpha &= 1.22 \\
\end{align*}
\]
(b) cut DA slack prior

\[ \beta \sim \text{Beta}(17.4, 4.6) \]

\[ \overbrace{\beta}^{d} \]

\[ \text{posterior} \]

\[ \beta \sim \text{Beta}(d+s, \beta + n-s) \]

\[ \overbrace{117.4}^{117.4} \]

\[ \overbrace{124.6}^{124.6} \]

\[ \begin{array}{|c|c|c|}
\hline
\text{prior} & \text{mean} & \text{SD} \\
\hline
\text{neutral} & 0.455 & 0.0333 \\
\text{cut+DA} & 0.455 & 0.0321 \\
\hline
\end{array} \]

\[ \text{posterior mean of } \theta \text{ is } \\
\frac{d+s}{d+\beta+n} \]

\[ \text{posterior SD is } \left( \frac{d+s}{d+\beta+n} \right) \left( \frac{\beta+n-s}{d+\beta+n} \right) \left( \frac{1}{d+\beta+n+1} \right) \]

The no-discrimination rate of 0.791 is

\[ \frac{0.791 - 0.455}{0.0333} = 10.6 \text{ standard deviations away from the posterior expectation} \]
under the neutral prior and

\[
\frac{0.791 - 0.485}{0.0321} = 9.5 \text{ posterior S.Ds}
\]

ey was just a lack prior; Q.E.D.

discrimination

Multinomial / You're contemplating a distribution that contains (back to discrete) elements of \( k \geq 2 \) types (e.g., \{Democrat, Republican, Libertarian, Independent, Green\}). Suppose the population of elements of type \( i \) is \( \theta_i \) with

\[
\sum_{i=1}^{k} \theta_i = 1; \quad \theta = (\theta_1, \ldots, \theta_k).
\]
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 \( \mathbf{X} = (X_1, \ldots, X_k) \) has

\[
\frac{M!}{p_1!p_2! \cdots p_k!} \left( \frac{1}{n} \right)^{\sum_{i=1}^{k} X_i} \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k}
\]

where

\[
\binom{n}{x_1, \ldots, x_k} = \frac{n!}{x_1!x_2! \cdots x_k!}
\]
is the multinomial coefficient.

This is called the multinomial \( (n, p) \) distribution.
\[ E(\xi_i) = n \mu_i \]
\[ \text{Var}(\xi_i) = np_i (1-p_i) \]
(just like binomial)

But now something new:

\[ C(\xi_i, \xi_j) = -np_i p_j \]

because \( \sum_{i=1}^{n} \xi_i = n \).

Bivariate Normal

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

\[ \xi_1, \xi_2 \overset{iid}{\sim} N(0, 1) \]

Specify 5 parameters:

<table>
<thead>
<tr>
<th>[ -\infty &lt; \mu_1 &lt; +\infty ]</th>
<th>[ 0 &lt; \sigma_1 &lt; \infty ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ -\infty &lt; \mu_2 &lt; +\infty ]</td>
<td>[ 0 &lt; \sigma_2 &lt; \infty ]</td>
</tr>
<tr>
<td>( -1 &lt; \rho &lt; 1 )</td>
<td></td>
</tr>
</tbody>
</table>
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 $\mathbf{X} = (X_1, X_2)$ is

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

This is the Bivariate Normal $(\mathbf{X}; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.
Easy to show that $E(X_1) = \mu_1$, $E(X_2) = \mu_2$, $V(X_1) = \sigma_1^2$, $V(X_2) = \sigma_2^2$,

$\rho(X_1, X_2) = \rho \cdot \sqrt{\text{Consequences of this Def.}}$.

1. $(X_1, X_2) \sim \text{Bivariate Normal} \rightarrow$

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

we already knew the $\rightarrow$ direction in general; what's new here is that

\text{correlation} \leftrightarrow \text{independence} if $(X_1, X_2) \sim \text{Bivariate Normal}$.\)
\( (X_1, X_2) \sim \text{Bivariate Normal} (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \)

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

\[ E(X_2 | x_1) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1) \]

\[ \text{variance } \text{Var}(X_2 | x_1) = (1 - \rho^2) \sigma_2^2 \]
And the means of all these normal distributions in the vertical strips are connected together by Galton's regression line:

\[
\hat{x}_2 = \mu_2 + \frac{\rho \sigma_2}{\sigma_1} (x_1 - \mu_1).
\]

This line has slope \( \frac{\rho \sigma_2}{\sigma_1} \) and \( \gamma \)-intercept \( \beta_0 = \mu_2 - \beta_1 \mu_1 \), hence:

\[
\hat{x}_2 = \beta_0 + \beta_1 x_1
\]

Moreover, we can now quantify an earlier insight:

- Ignore \( x_1 \), predict \( \hat{x}_2 \mid x_1 = \mu_2 = E(x_2 \mid x_1) \)

The root mean squared (RMS) of this prediction is

\[
\sqrt{\text{V}(\hat{x}_2)} = \sigma_2
\]
Use $x_1$ to predict $x_2$

$$\text{pred.2} \left( x_2 \right)_{x_1} = E \left( x_2 | x_1 = x_1 \right)$$

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

Part of the prediction is:

$$\sqrt{V \left( x_2 | x_1 \right)} = \sigma_2 \sqrt{1 - \rho^2}$$

Since $\rho < 1$, \( \sigma_2 \sqrt{1 - \rho^2} \leq \sigma_2 \) with equality only when $\rho = 0$.

\( \Rightarrow \) $(x_1, x_2) \sim \text{Bivariate Normal} (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$

$z = a_1 x_1 + a_2 x_2 + b$, $(a_1, a_2, b)$ arbitrary constants

$\rightarrow z \sim N \left( a_1 \mu_1 + a_2 \mu_2 + b, a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + 2a_1 a_2 \rho \sigma_1 \sigma_2 \right)$. 
You draw an IID random sample $X_1, \ldots, X_n$ 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 $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ is the best you can do (in the absence of prior information). It would be nice if $\overline{X}_n$ approached the right answer $\mu$ as $n$ increases; how to quantify that idea?
You can't move more of the symbol right tail beyond a certain point. 

Again, the symbol is fixed, work on stochastic processes.

\[ P(T > e) = \frac{1}{e} \quad \text{for all } e > 0 \]

\[ \sup \lim E(X) \geq \frac{1}{e} \quad \text{for all } e > 0 \]

Markov

\[ 0 \leq e \leq 1 \]

Two

Markov inequalities

\[ P(\mathbb{E}[X] < 0) = \text{help} \]

\[ \mathbb{P}(X \geq 0) = 1 \]

\[ \mathbb{E}[X] \geq 0 \]

\[ \sup \lim E(X) \geq \frac{1}{e} \quad \text{for all } e > 0 \]

\[ \sup \lim E(X) \geq \frac{1}{e} \quad \text{for all } e > 0 \]

\[ \sup \lim E(X) \geq \frac{1}{e} \quad \text{for all } e > 0 \]