96. **Beta distribution** is great for modeling the mean (p) of a Bernoulli distribution. Beta(10, 10) indicates 10 successes in 20 trials, and Beta(100, 100) indicates 100 successes in 200 trials. The two distributions have different shapes where the latter is thinner since the credibility is higher in that experiment.

96. The reason why we assume the reciprocal of variance follows Gamma distribution, is that the sum of squared standard normal random variables follows chi-square, which is a special case of Gamma distribution.

97. **Bayesian stats conjugate family**: given likelihood function P(x|theta), if the prior P(theta) and posterior P(theta|x) follow the same family of distribution, then we say the likelihood and the prior are conjugate.

e.g. Binomial likelihood and Beta prior, Poisson likelihood and Gamma prior.

98. To counter the criticism that prior elicitation is hard and subjective, google **Jeffrey's prior**. There's a whole science behind how to infer the prior distribution. Most recently, reference prior is the state-of-the-art. Basically, it determines the prior by maximizing the divergence between the prior and the posterior. Jeffery's prior and reference prior coincide in 1d case.

99. Just remember the terminology: posterior is likelihood times prior divided by marginal. Frequentists will focus on the likelihood and maybe try to maximize it. posterior is proportional to the joint.

100. There's nothing mystical about generalized linear regression now. The model you fit is essentially the (conditional) mean parameter of the distribution of the Y variable.

**Linear regression** is essentially Y ~ Normal(B0 + B1 \* x1 + B2 \* x2, eps ^ 2) where (1 / eps ^ 2) is usually assumed to have Gamma distribution with small alpha and beta, e.g., 1e-3, 1e-3 if for **noninformative** priors, and B0, B1, B2 are assumed to have Normal distributions with large variance e.g., 1e5 if for **noninformative**priors.

**Poisson regression** is essentially Y ~ Poisson(lambda = B0 + B1 \* x1 + B2 \* x2) where B0, B1, B2 are again typically Gaussian distributions.

**Logistic regression** is essentially Y ~ Bernoulli(p = 1 / (1 + exp[-(B0 + B1 \* x1 + B2 \* x2) ]))

**Binomial regression** is essentially Y ~ Binomial(p = 1 / (1 + exp[-(B0 + B1 \* x1 + B2 \* x2) ]), n = Nobs)

Check Wikipedia Table "**Common distributions with typical uses and canonical link functions**" at <https://en.wikipedia.org/wiki/Generalized_linear_model> . The reason behind those parametric distribution forms is Bayesian inference for the coefficient estimates. If given no other information, the priors for the betas are iid Gaussian with 0 mean and very large variance. Check <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>

for uninformative prior. With the prior distributions, regression becomes Bayesian regression. Do not confuse it with Gaussian process regression.

**Binomial regression and logistic regression** are almost identical in nature. Consider data like:

success  trials  X1  X2

10          20     2    3

5            6       7    8

...

Such data in appearance calls for binomial regression, but in fact it can be disaggregated into data where trials is absent, and success is binary, thus logistic regression is called

**101. Metropolis Hastings**: it is basically an exploration algorithm where the agent tends to move to areas with higher probabilities. Given a proportional probability density function (usually in an analytical form), we take an initial point and add some random noise to it. If the new point has higher probability density, we accept it into the sample, otherwise we reject it with a certain probability, and make a copy of the old point as the new point. Then we add some random noise to the new point and repeat. With some mild assumptions about the random noise and about the probability density function, we can prove the sample distribution will converge to the theoretical.

See Wikipedia for how to compute the acceptance probability. Start with thinking how to sample a Bernoulli distribution to have an intuition of why the acceptance probability is computed that way.

Metropolis Hastings apparently can be used to find the minimum/maximum of any function. This gives rise to the simulated annealing algorithm.

**102. Gibbs sampler**: it is just a special implementation of Metropolis Hastings (MH). It is used when the PDF is high dimensional and computationally expensive, while the conditional PDFs are cheap. For example, consider sampling from f(X1, X2, X3) that is hard to compute. Given initial points x1, x2, we can (i) sample an x3 from f(X3|X1 = x1, X2 = x2) via MH, (ii) sample a new x1 from f(X1|X2 = x2, X3 = x3) via MH, (iii) sample a new x2 from f(X2|X1 = x1, X3 = x3) and go back to (i).

In the sampled sequence generated by MH, elements close to each other are also often close in space, or “correlated”. This is because (i) MH copies the old point as the new point in the case of rejection, and (ii) the random noise added during exploration may not have a large variance.

The **burn-in** period refers to the initial segment of samples in the sequence produced by MH. This segment should be ignored after MH converges.

**Hamilton Monte Carlo and NUTS (No U-turn sampler)**: this requires computing the gradient of the PDF. The agent follows the gradient and jumps to a section where probabilities are likely high.

**103. Bayesian GLM / pymc:**

# ==============================================================================  
# Nonlinear regression model: y = exp(sin(b0 + b1 \* x1) + tan(b2 \* x2))  
# In Bayesian context:  
#   Y ~ Normal( exp(sin(B0 + B1 \* x1) + tan(B2 \* x2)), 100)  
# where (for example) B0 ~ Normal(0, 1), B1 ~ Gamma(1, 2), B2 ~ t(2),   
# B0, B1, B2 are independent (assumption), 100 as Y's standard deviation (non-informative prior assumption).  
# The first misunderstanding in Bayesian GLM is the dreadful thought of  
# finding the distribution of the entity `exp(sin(B0 + B1 \* x1) + tan(B2 \* x2))`.  
# Notice this entity is not Y, but merely the mean parameter of the likelihood  
# distribution.  
# ==============================================================================  
set.seed(123)  
x1 = round(rnorm(5), 2)  
x2 = round(rnorm(5), 2)  
yobserved = round(rexp(5), 2)  
loglikelihood = sum(dnorm(  
  y, mu = exp( sin(b0 + b1 \* x1) + tan(b2 \* x2) ), sd = 100, log = T))  
  
# Likelihood: probability of the observed y given b0, b1, b2:  
loglikelihood = function(b0, b1, b2)   
{  
  sum(dnorm(y, mu = exp( sin(b0 + b1 \* x1) + tan(b2 \* x2) ),   
            sd = 1, log = T))  
}  
  
# Prior: probability of b0, b1, b2.  
logPrior = function(b0, b1, b2)  
{  
  dnorm(b0, 0, 1, log = T) + dgamma(b1, shape = 1, rate = 2, log = T) +  
    dt(b2, 2, log = T)  
}  
  
# The joint or the proportional posterior.   
logJoint = function(b0, b1, b2)  
{  
  loglikelihood(b0, b1, b2) + logPrior(b0, b1, b2)  
}  
# Now one can sample the posterior using MH or Gibbs. For Gibbs, although

# here we have the joint, but (b0|b1, b2) is just proportional to (b0, b1, b2) when

# b1 and b2 are fixed.  
# ==============================================================================

104. Idea for **nonparametric prior distribution:** given a sample, create a histogram out of it. The bin width is chosen by the Freedman-Diaconis rule: width = 2 \* (Q(75%) - Q(25%)) \* N ^ (-1 / 3) where N is the sample size. Use half of the SD of the sample as the SD of the random noise in Metropolis Hastings.

105. Consider random sampling with replacement from N objects. What's the minimum sample size to ensure the probability of all objects being sampled is greater than p? The answer is harder than it looks and involves sterling number of the second kind: <https://math.stackexchange.com/questions/379525/probability-distribution-in-the-coupon-collectors-problem>

106. **Frequentist optimization for the parameter**vs. **Bayesian posterior mean of the parameter**: With noninformative priors, the reason why these two can be similar is because the mean is often equal or close to the mode of a distribution.