Bayesian (and Frequentist) Principles

Ben Goodrich

February 21, 2022

Predictive Distributions

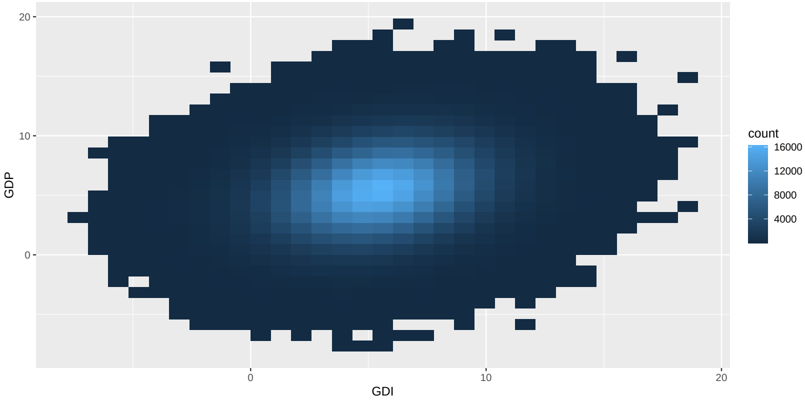
- It is at best difficult and usually impossible to derive the denominator of Bayes Rule before seeing the data, e.g. $f\left(\cancel{\mu} \bigcap I \bigcap P \mid m, s, \sigma_I, \sigma_P, \rho\right) =$

$$\int_{-\infty}^{\infty} f\left(\mu \mid m, s\right) f\left(I \mid \mu, \sigma_I\right) f\left(P \mid \mu + \frac{\sigma_P}{\sigma_I} \rho\left(I - \mu\right), \sigma\sqrt{1 - \rho^2}\right) d\mu$$

- But it is easy to draw from this predictive distribution of reported GDI & GDP

```
library(ggplot2)
library(dplyr)
m <- 5.45; s <- 1.5; sigma <- 7 / 3; rho <- -1 / 10
# together these are a trivariate random variable; separat
tibble(mu  = rnorm(10^6, mean = m,  sd = s),
       GDI = rnorm(10^6, mean = mu, sd = sigma),
       GDP = rnorm(10^6, mean = mu + rho * (GDI - mu),
                   sd = sigma * sqrt(1 - rho^2))) %>%
  ggplot() + geom_bin_2d(aes(x = GDI, y = GDP)) # plot on
```

Plot from Previous Slide



Prior Predictive Probability for Bowling

$f\left(\theta \bigcap x_1 \bigcap x_2 \mid m, n = 10\right) =$
$\int_0^{\infty} f\left(\theta \mid m\right) \Pr\left(x_1 \mid n = 10, \theta\right) \Pr\left(x_2 \mid n = 10 - x_1, \theta\right) d\theta =$
$\Pr\left(x_1 \bigcap x_2 \mid m, n = 10\right)$, but that area can only be calculated by calling `integrate`

	0	
0	0.0004363	
1	0.0005431	
2	0.0006928	
3	0.0009111	
4	0.0012459	
5	0.0017952	
6	0.0027824	
7	0.0048108	
8	0.009985	
9	0.0301002	
10	0.5096771	<

Simulated Prior Predictive Probability for Bowling

```
tibble(theta = rexp(10^4, rate = 1 / 0.15),
       x_1 = sapply(theta, FUN = function(t) {
         sample(Omega, size = 1, prob = Pr(Omega, n = 10, t
       })) %>%
  group_by(x_1) %>%
  mutate(x_2 = sapply(theta, FUN = function(t) {
    sample(Omega, size = 1, prob = Pr(Omega, n = 10 - first
  })) %>%
  ungroup %>%
  with(., table(x_1, x_2)) %>%
  prop.table %>%
  round(digits = 6)
```

```
##      x_2
## x_1       0      1      2      3      4      5      6
##   0  0.0001 0.0005 0.0005 0.0004 0.0005 0.0010 0.0011 0.
##   1  0.0012 0.0005 0.0010 0.0009 0.0008 0.0011 0.0014 0.
##   2  0.0008 0.0005 0.0009 0.0007 0.0012 0.0012 0.0022 0.
```

Prior Predictive Distribution for Future Data

- Before the data are observed, i.e. on HW2, the denominator of Bayes Rule

$$f\left(\cancel{\theta}\bigcap \mathbf{y} \mid \ldots\right) = \int_{\Theta} f\left(\theta \mid \ldots\right) f\left(\mathbf{y} \mid \theta\right) d\theta$$

  defines a "prior" P{D,M}F for $\mathbf{y}$
- One way to tell whether your prior distribution for the PARAMETERS, θ, is reasonable is to judge whether the implied prior predictive distribution for the OUTCOMES is reasonable
- For example, is a prior probability of a strike that is about $\frac{1}{2}$ reasonable for a woman competing at the World Cup of Bowling? Is it reasonable for a man?

- You can draw from the POSTERIOR predictive distribution of future data by repeatedly drawing θ from is posterior distribution given past data and using those realizations to draw $\widetilde{\mathbf{y}}$ from its conditional distribution given θ

Ex Ante Probability (Density) of *Ex Post* Data

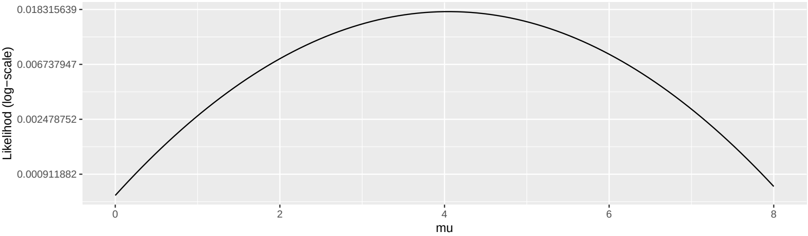
A likelihood function is the same expression as a P{D,M}F with 3 distinctions:

1. For the PDF or PMF, $f(x|\boldsymbol{\theta})$, we think of X as a random variable and $\boldsymbol{\theta}$ as given, whereas we conceive of the likelihood function, $\mathcal{L}(\boldsymbol{\theta}; x)$, to be a function of $\boldsymbol{\theta}$ (in the mathematical sense) evaluted at the OBSERVED data, x
   - As a consequence, $\int\limits_{-\infty}^{\infty} f(x|\boldsymbol{\theta})\, dx = 1$ or $\sum\limits_{x\in\Omega} f(x|\boldsymbol{\theta}) = 1$ while $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \mathcal{L}(\boldsymbol{\theta}; x)\, d\theta_1 d\theta_2 \dots d\theta_K$ may not exist and is never 1
2. We often think of "the likelihood function" for N conditionally independent observations, so $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \prod_{n=1}^{N} \mathcal{L}(\boldsymbol{\theta}; x_n)$
3. By "the likelihood function", we often really mean the natural logarithm thereof, a.k.a. the log-likelihood function $\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln \mathcal{L}(\boldsymbol{\theta}, \mathbf{x}) = \sum_{n=1}^{N} \ln \mathcal{L}(\boldsymbol{\theta}; x_n)$

Maximum Likelihood Estimation (MLE)

- What is the MLE for μ (often denoted $\widehat{\mu}$) in the third quarter of 2021, when GDI growth was 5.8 and GDP growth was 2.3?

```
ggplot() + xlim(0,8) + ylab("Likelihod (log-scale)") + xlab
  scale_y_continuous(trans = "log") +
  geom_function(fun = ~dnorm(5.8, mean = .x, sd = sigma) *
                       dnorm(2.3, mean = .x + rho * (5.8 -
```



Subtlties of Maximum Likelihood Estimation

- $\widehat{\mu}$ is NOT the most likely value of μ given that GDI growth was 5.8 and GDP growth was 2.3 because μ is not a random variable and Frequentist probability does not apply to it (just like Cook's huge odd integer)
- $\widehat{\mu}$ IS the value of μ such that the most likely values of the random variables GDI growth and GDP growth are 5.8 and 2.3 respectively (so MLEs overfit)
- Could other values of μ yield GDI growth of 5.8 and GDP growth of 2.3? Yes.
- Since $\int_{-\infty}^{\infty} L(\mu)\, d\mu \neq 1$, the likelihood function is in arbitrary units (not density), but that does not affect the maximization of it
- Instead of maximizing $L(\mu)$, Bayesians divide $L(\mu)$ by $f(\cancel{\mu} \bigcap I \bigcap P \mid \dots)$, which is in the same arbitrary units. Thus, the arbitrary units cancel leaving the posterior PDF in the same units as the prior PDF, which both integrate to 1 over all possible values of μ.

Probability Distribution of the MLE

- For Frequentists, $\widehat{\mu} = \arg\max L(\mu; \mathbf{y})$ is a random variable whose distribution is conditioned on μ (or whatever the parameters are), with PDF

$$f(\widehat{\mu} \mid \mu) = f\left(\widehat{\mu} \bigcap \cancel{\mathbf{y}} \mid \mu\right) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\mu}(\mathbf{y})\, \underset{\text{multivariate}}{f(\mathbf{y} \mid \mu)}\, dy_1 \dots dy_N$$

- Thus, the probability distribution of $\widehat{\mu} \mid \mu$ is irrespective of the data but weighted by the PDF $f(\mathbf{y} \mid \mu)$, which is completely different from Bayesians' prior predictive density but obtained in a similar way

$$f(\mathbf{y} \mid \dots) = f\left(\cancel{\theta} \bigcap \mathbf{y} \mid \dots\right) = \int_{-\infty}^{\infty} f(\theta \mid \dots)\, f(\mathbf{y} \mid \theta)\, d\theta$$

- As $N \uparrow \infty$, $f(\widehat{\mu} \mid \mu) \to \frac{e^{-\frac{1}{2\nu}\left(\widehat{\mu} - \mu\right)^2}}{\sqrt{2\pi\nu}}$, which is the PDF of a normal distribution with expectation μ and variance $\nu \propto \frac{1}{N}$, across random samples of size N

Frequentist Analysis with Moderna Trial Data

- 11 fully vaccinated people got covid and 185 placeboed people got covid

```
theta_0 <- (0.3 - 1) / (0.3 - 2) # 0.412, implied by a hyp
binom.test(c(11, 185), p = theta_0, alternative = "less")
```

```
##
##  Exact binomial test
##
## data:  c(11, 185)
## number of successes = 11, number of trials = 196, p-valu
## alternative hypothesis: true probability of success is l
## 95 percent confidence interval:
##  0.00000000 0.09118634
## sample estimates:
## probability of success
##             0.05612245
```

```
CI_theta_high <- 0.09118634
CI_VE_low <- (1 - 2 * CI_theta_high) / (1 - CI_theta_high)
```

What Is Going On with binom.test?

```
n <- 11 + 185
p_value <- pbinom(11, size = n, prob = theta_0)          #
CI_theta_high <- qbeta(0.95, shape1 = 11 + 1, shape2 = n) #
```

```
library(dplyr)
tibble(y = rbinom(10^6,  size = n, prob = theta_0), # y is
       p_val = pbinom(y, size = n, prob = theta_0), # rando
       CI_high = qbeta(0.95, shape = y + 1, shape2 = n - y)
  summarize(type_I_error = mean(p_val < 0.05),     # simula
            catch_rate = mean(theta_0 < CI_high)) # simula
```

```
## # A tibble: 1 x 2
##   type_I_error catch_rate
##          <dbl>      <dbl>
## 1       0.0371      0.963
```

- In general, type_I_error is at most 0.05 and catch_rate is at least 0.95, but for continuous test statistics these bounds are usually tight

Applied Statistics Is Not Actually Frequentist

- It is objectively, provably true that `pbinom(11, size = n, prob = theta_0)` is the right cumulative probability under its (possibly strong) assumptions frequentist but not relevant!!

Applied Statistics Is Not Actually Frequentist

- It is objectively, provably true that `pbinom(11, size = n, prob = theta_0)` is the right cumulative probability under its (possibly strong) assumptions
- That only becomes relevant under the CONVENTION that we reject the null hypothesis that $\theta = \theta_0$ if and only if that cumulative probability is < 0.05

Applied Statistics Is Not Actually Frequentist

- It is objectively, provably true that `pbinom(11, size = n, prob = theta_0)` is the right cumulative probability under its (possibly strong) assumptions
- That only becomes relevant under the CONVENTION that we reject the null hypothesis that $\theta = \theta_0$ if and only if that cumulative probability is < 0.05
- Deciding that $\theta \neq \theta_0$ or even $\theta < \theta_0$ is insufficient for science (or anything else), which needs to know something about what θ is rather than what it isn't

Applied Statistics Is Not Actually Frequentist

- It is objectively, provably true that `pbinom(11, size = n, prob = theta_0)` is the right cumulative probability under its (possibly strong) assumptions
- That only becomes relevant under the CONVENTION that we reject the null hypothesis that $\theta = \theta_0$ if and only if that cumulative probability is < 0.05
- Deciding that $\theta \neq \theta_0$ or even $\theta < \theta_0$ is insufficient for science (or anything else), which needs to know something about what θ is rather than what it isn't
- That only becomes relevant under the CONVENTION that if and only if you reject the null hypothesis that $\theta = \theta_0$, then you are "allowed" to proceed as if $\theta = \widehat{\theta}$ or to appear more sophisticated, proceed as if $\theta \sim \mathcal{N}\left(\widehat{\theta}, SE\left(\widehat{\theta}\right)\right)$. If you fail to reject the null hypothesis that $\theta = \theta_0$, then proceed as if $\theta = \theta_0$

Applied Statistics Is Not Actually Frequentist

- It is objectively, provably true that `pbinom(11, size = n, prob = theta_0)` is the right cumulative probability under its (possibly strong) assumptions
- That only becomes relevant under the CONVENTION that we reject the null hypothesis that $\theta = \theta_0$ if and only if that cumulative probability is < 0.05
- Deciding that $\theta \neq \theta_0$ or even $\theta < \theta_0$ is insufficient for science (or anything else), which needs to know something about what θ is rather than what it isn't
- That only becomes relevant under the CONVENTION that if and only if you reject the null hypothesis that $\theta = \theta_0$, then you are "allowed" to proceed as if $\theta = \widehat{\theta}$ or to appear more sophisticated, proceed as if $\theta \sim \mathcal{N}\left(\widehat{\theta}, SE\left(\widehat{\theta}\right)\right)$. If you fail to reject the null hypothesis that $\theta = \theta_0$, then proceed as if $\theta = \theta_0$
- None of those conventions are Frequentist, objective, or even good ideas

More on Confidence Intervals

- Jerzy Neyman, who invented the confidence interval, said *I have repeatedly stated that the frequency of correct results will tend to α [the type I error rate]. Consider now the case when a sample is already drawn, and the calculations have given [particular limits]. Can we say that in this particular case the probability of the true value [falling between these limits] is equal to α? The answer is obviously in the negative. The parameter is an unknown constant, and no probability statement concerning its value may be made . . . "*

- So, the FDA's rule that in order to approve a vaccine — the confidence interval for the VE should exclude 0.3 — does not imply there is a 0.95 probability that the VE is greater than 0.3 and contradicts how the creator of confidence intervals says they should be used

More on Confidence Intervals

- Jerzy Neyman, who invented the confidence interval, said *I have repeatedly stated that the frequency of correct results will tend to α [the type I error rate]. Consider now the case when a sample is already drawn, and the calculations have given [particular limits]. Can we say that in this particular case the probability of the true value [falling between these limits] is equal to α? The answer is obviously in the negative. The parameter is an unknown constant, and no probability statement concerning its value may be made . . . "*
- So, the FDA's rule that in order to approve a vaccine — the confidence interval for the VE should exclude 0.3 — does not imply there is a 0.95 probability that the VE is greater than 0.3 and contradicts how the creator of confidence intervals says they should be used
- A confidence interval is a range of values such that if the null hypothesis value, θ_0, were anywhere in that interval, you

Tips to Avoid Being Confused by Frequentism

- What is the nature and timing of the randomization? (Often non-existent)

Tips to Avoid Being Confused by Frequentism

- What is the nature and timing of the randomization? (Often non-existent)
- Probability looks forward from the instant before the randomization and conditions on everything previous, including θ, which is weird

Tips to Avoid Being Confused by Frequentism

- What is the nature and timing of the randomization? (Often non-existent)
- Probability looks forward from the instant before the randomization and conditions on everything previous, including $\boldsymbol{\theta}$, which is weird
- Only make probability statements about random variables, such as data, estimators, and test statistics. Do not make probability statements about constants you conditioned on, such as $\boldsymbol{\theta}$, hypotheses, and research designs.

Tips to Avoid Being Confused by Frequentism

- What is the nature and timing of the randomization? (Often non-existent)
- Probability looks forward from the instant before the randomization and conditions on everything previous, including $\boldsymbol{\theta}$, which is weird
- Only make probability statements about random variables, such as data, estimators, and test statistics. Do not make probability statements about constants you conditioned on, such as $\boldsymbol{\theta}$, hypotheses, and research designs.
- Instead of saying "the probability of A", say "the proportion of times that A would happen over the (hypothetical) randomizations of ..."

Tips to Avoid Being Confused by Frequentism

- What is the nature and timing of the randomization? (Often non-existent)
- Probability looks forward from the instant before the randomization and conditions on everything previous, including $\boldsymbol{\theta}$, which is weird
- Only make probability statements about random variables, such as data, estimators, and test statistics. Do not make probability statements about constants you conditioned on, such as $\boldsymbol{\theta}$, hypotheses, and research designs.
- Instead of saying "the probability of A", say "the proportion of times that A would happen over the (hypothetical) randomizations of ..."
- Instead of saying some estimator is consistent, unbiased, efficient, etc., insert the definitions. E.g., "The average of $\widehat{\theta}$ across random sampled datasets of fixed size N is θ (unbiased)" or "As $N \uparrow \infty$, the average squared difference between θ and $\widehat{\theta}$ across random sampled datasets diminishes (consistent)".

Tips to Avoid Being Confused by Frequentism

- What is the nature and timing of the randomization? (Often non-existent)
- Probability looks forward from the instant before the randomization and conditions on everything previous, including $\boldsymbol{\theta}$, which is weird
- Only make probability statements about random variables, such as data, estimators, and test statistics. Do not make probability statements about constants you conditioned on, such as $\boldsymbol{\theta}$, hypotheses, and research designs.
- Instead of saying "the probability of A", say "the proportion of times that A would happen over the (hypothetical) randomizations of ..."
- Instead of saying some estimator is consistent, unbiased, efficient, etc., insert the definitions. E.g., "The average of $\widehat{\theta}$ across random sampled datasets of fixed size N is θ (unbiased)" or "As $N \uparrow \infty$, the average squared difference between θ and $\widehat{\theta}$ across random sampled datasets diminishes (consistent)".

Four Ways to Execute Bayes Rule

1. Utilize conjugacy to analytically integrate the kernel of Bayes Rule
   - Makes incremental Bayesian learning obvious but is only possible in simple models when the distribution of the outcome is in the exponential family
2. Numerically integrate the kernel of Bayes Rule over the parameter(s) only feasible with 1 or 2 dimensions of unknown
   - Most similar to what we did in the discrete case but is only feasible when there are very few parameters and can be inaccurate even with only one
3. Draw from the joint distribution and keep realizations of the parameters if and only if the realization of the outcome matches the observed data only for discrete
   - Very intuitive what is happening but is only possible with discrete outcomes and only feasible with few observations and parameters
4. Perform MCMC (via Stan) to randomly draw from the posterior distribution
   - Works for any posterior PDF that is differentiable w.r.t. the

(1) Conjugacy

- Prior is a Beta distribution with shape parameters $a > 0$ and $b > 0$, which has the PDF

$$f(\theta \mid a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 t^{a-1}(1-t)^{b-1}\,dt}$$

- Likelihood, $L(\theta; y, n)$ is binomial with y successes in n tries, which has the same form, ${n \choose y}\theta^y(1-\theta)^{n-y}$, as the prior PDF
- Posterior is a Beta distribution with shape parameters $a^* = a + y$ and $b^* = b + n - y$ and normalizing constant $\frac{1}{B(a^*,b^*)}$
- For Biontech / Pfizer from Week04,

```
a <- 0.700102; b <- 1; y <- 8; n <- 94; a_star <- a + y; b_
```

(2) Numerical Integration

- Suppose you did not known the beta prior and binomial likelihood were naturally conjugate
- You could still calculate the area in the denominator numerically to obtain the posterior PDF, which works in general (with 1 parameter)

```
kernel <- function(theta) { # omits constant choose(n, y) /
  theta^(a - 1) * (1 - theta)^(b - 1) * theta^y * (1 - thet
}
constant <- integrate(kernel, lower = 0, upper = 1)$value #
format(rbind(exactish = dbeta(0.1, a_star, b_star),
             approximate = kernel(0.1) / constant), digits
```

```
##               [,1]
## exactish    "11.8010273"
## approximate "11.8010268"
```

(3) Filter the Joint Distribution using (Discrete) y

```
filtered <- tibble(theta = rbeta(10^6, a, b)) %>%
  filter(y == rbinom(n(), size = n, prob = theta))
rbind(exact = choose(n, y) * beta(a_star, b_star) / beta(a,
      simulated = nrow(filtered) / 10^6)
```

```
##                 [,1]
## exact     0.01529499
## simulated 0.01531300
```

```
rbind(exact = a_star / (a_star + b_star),
      simulated = mean(filtered$theta))
```

```
##                 [,1]
## exact     0.09091006
## simulated 0.09146953
```

(4) Specify a Posterior in Stan to Draw from

```
data {
  int<lower  = 0> n;              // tries
  int<lower  = 0, upper = n> y; // successes

  real<lower = 0> a; // a and b are knowns
  real<lower = 0> b; // so they count as data
}
parameters {
  real<lower = 0, upper = 1> theta; // success probability
}
model {
  // _lp{d,m}f means "logarithm of P{D,M}F" for numerical r
  target += beta_lpdf(theta | a, b) + binomial_lpmf(y | n,
} // denominator of Bayes Rule is not necessary or utilizec
generated quantities { // of interest, but not part of the
  real VE = (1 - 2 * theta) / (1 - theta);
}
```

(4) Resulting MCMC Output

```
post <- rstan::read_stan_csv("post.csv")
post
```

```
## Inference for Stan model: post.
## 1 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draw
##
##        mean se_mean   sd  2.5%   25%   50%   75% 97.5% n
## theta  0.09    0.00 0.03  0.04  0.07  0.09  0.11  0.16
## VE     0.90    0.00 0.04  0.81  0.88  0.90  0.93  0.96
## lp__  -4.65    0.04 0.77 -6.73 -4.83 -4.35 -4.14 -4.08
##
## Samples were drawn using NUTS(diag_e) at Mon Feb 21 4:10
## For each parameter, n_eff is a crude measure of effectiv
## and Rhat is the potential scale reduction factor on spli
## convergence, Rhat=1).
```

A Better (but incomplete) Stan Program

```
data {
  int<lower  = 0> n;                // tries
  int<lower  = 0, upper = n> y; // successes

  // more hyperparameters for the prior on VE
}
parameters {
  real<upper = 1> VE;
}
transformed parameters {
  real theta = (VE - 1) / (VE - 2); // implied success prob
}
model {
  target += binomial_lpmf(y | n, theta);
  target += // some prior on VE
}
```

Principles to Choose Priors With

1. Do not use improper priors (those that do not integrate to 1)
2. Subjective, including "weakly informative" priors
3. Entropy Maximization
4. Invariance to reparameterization (particularly scaling)
5. "Objective" (actually also subjective, but different from 2)
6. Penalized Complexity (PC) (which we will cover the last week of the semester)

- Choose a prior family that integrates to 1 over the parameter space, Θ

Principles to Choose Priors With

1. Do not use improper priors (those that do not integrate to 1)
2. Subjective, including "weakly informative" priors
3. Entropy Maximization
4. Invariance to reparameterization (particularly scaling)
5. "Objective" (actually also subjective, but different from 2)
6. Penalized Complexity (PC) (which we will cover the last week of the semester)

- Choose a prior family that integrates to 1 over the parameter space, Θ
- Then choose hyperparameter values that are consistent with what you believe

Principles to Choose Priors With

1. Do not use improper priors (those that do not integrate to 1)
2. Subjective, including "weakly informative" priors
3. Entropy Maximization
4. Invariance to reparameterization (particularly scaling)
5. "Objective" (actually also subjective, but different from 2)
6. Penalized Complexity (PC) (which we will cover the last week of the semester)

- Choose a prior family that integrates to 1 over the parameter space, Θ
- Then choose hyperparameter values that are consistent with what you believe
- The important part of a prior is what values it puts negligible probability on

Principles to Choose Priors With

1. Do not use improper priors (those that do not integrate to 1)
2. Subjective, including “weakly informative” priors
3. Entropy Maximization
4. Invariance to reparameterization (particularly scaling)
5. “Objective” (actually also subjective, but different from 2)
6. Penalized Complexity (PC) (which we will cover the last week of the semester)

- Choose a prior family that integrates to 1 over the parameter space, Θ
- Then choose hyperparameter values that are consistent with what you believe
- The important part of a prior is what values it puts negligible probability on
- Draw from the prior predictive distribution of $\mathbf{y}$ to see if it makes sense

Dirichlet Distribution

- Dirichlet distribution is over the parameter space of PMFs — i.e. $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$ — and the Dirichlet PDF is $f\left(\boldsymbol{\pi} \mid \boldsymbol{\alpha}\right) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$ where $\alpha_k \geq 0 \, \forall k$ and the multivariate Beta function is $B\left(\boldsymbol{\alpha}\right) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\prod_{k=1}^{K} \alpha_k\right)}$ where $\Gamma\left(z\right) = \frac{1}{z} \prod_{n=1}^{\infty} \frac{\left(1+\frac{1}{n}\right)^n}{1+\frac{z}{n}} = \int_0^{\infty} u^{z-1} e^{-u} du$ is the Gamma function
- $\mathbb{E}\pi_i = \frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k} \forall i$ and the mode of π_i is $\frac{\alpha_i - 1}{-1 + \sum_{k=1}^{K} \alpha_k}$ if $\alpha_i > 1$
- Iff $\alpha_k = 1 \, \forall k$, $f\left(\boldsymbol{\pi} \middle| \, \boldsymbol{\alpha} = \mathbf{1}\right)$ is constant over Θ (simplexes)
- Beta distribution is a special case of the Dirichlet where $K = 2$
- Marginal and conditional distributions for subsets of $\boldsymbol{\pi}$ are also Dirichlet
- Dirichlet distribution is conjugate with the multinomial and categorical

Multinomial Distribution

- The multinomial distribution over $\Omega = \{0, 1, \ldots, n\}$ has a PMF $\Pr\left(x \middle| \pi_1, \pi_2, \ldots, \pi_K\right) = n! \prod_{k=1}^{K} \frac{\pi_k^{x_k}}{x_k!}$ where the parameters satisfy $\pi_k \geq 0 \forall k$, $\sum_{k=1}^{K} \pi_k = 1$, and $n = \sum_{k=1}^{K} x_k$
- The multinomial distribution is a generalization of the binomial distribution to the case that there are K possibilities rather than merely failure vs. success
- Categorical is a special case where $n = 1$
- The multinomial distribution is the count of n independent categorical random variables with the same π_k values
- Draw via `rmultinom(1, size = n, prob = c(pi_1, pi_2, ..., pi_K))`