# PhD Bayesian Analysis:
# Session Three: Priors

INSEAD

The Business School
for the World®

# Revisit: Tossing a coin

- I tossed a coin 12 times, with the following results:

$$H, T, H, H, H, H, H, T, H, H, H, T$$

$$9H, 3T$$

- Want to predict the probability that the coin will land heads on the next (13th) toss
- Karl Pearson considers this an important problem in classical statistics.
- In Bayesian statistics, it motivates Laplace's well-known Rule of Succession (which we will learn about today).
- Question is: how do you start to make this inference? What else do you want to know from me?

# Classical approach

- Let's perform a simple hypothesis test of $H_0 : \theta = 0.5$.
- Need to compute the probability of observing our result, or a more extreme result, under $H_0$.

$$
\begin{aligned}
P(X \geq 9 | \theta = 0.5) &= \sum_{x=9}^{12} \binom{12}{x_i} \left(\frac{1}{2}\right)^{x_i} \left(\frac{1}{2}\right)^{12-x_i} \\
&= \left[\binom{12}{3} + \binom{12}{2} + \binom{12}{1} + \binom{12}{0}\right] \left(\frac{1}{2}\right)^{12} \\
&= \frac{299}{4096} \approx 0.07
\end{aligned}
$$

- The result is not statistically significant. Cannot reject the hypothesis that coin is fair.

# Classical approach



- But wait! Who said that more extreme than 9, 3 would be 10, 2; 11, 1; etc.
- What if I was tossing the coin until I get 3 tails? Than more extreme would be 10, 3; 11, 3; 12, 3; and so on. Under this assumption, the $p$-value is as follows:

$$
\begin{aligned}
P(Y \geq 12|\theta = 0.5) &= \sum_{x=12}^{\infty} \binom{x_i - 1}{r - 1} .5^r .5^{x_i - r} \\
&= \sum_{x=12}^{\infty} \binom{x_i - 1}{2} .5^{x_i} \\
&= 1 - \sum_{x=1}^{11} \binom{x_i - 1}{2} .5^{x_i} \\
&\approx 0.03
\end{aligned}
$$

- Now the result *is* statistically significant!

# Classical approach



- Does it make sense that my choice of when to stop collecting data affects the inferences you can draw about the coin? Why do you care? The data is what it is.
- What if I stopped because I wanted to get a sandwich?
- Eg. HIV antiretroviral drug example (ethical reasons to stop collecting data).
- Using language we learned earlier: this is a case where classical statistics violates the likelihood principle because the likelihoods are proportional to $\theta^x(1-\theta)^{n-x}$ but the inferences are different.

# Exchangeability

- Notice that, above, we assumed the data generating process is $iid$. This is a typcial assumption in classical inference.
- In Bayesian inference, we will assume something weaker, *exchangeability*.
- The sequence of tosses is exchangeable if $p(x_1, ..., x_n) = p(x_{\sigma(1)}, ..., x_{\sigma(n)})$ for every permutation $\sigma$ of the order.
- Position and order is irrelevant, for any length of the sequence, not just $12$.
- Exchangeability is an assumption about the underlying symmetry in the inference problem.
- Independent Bernoulli trials with $x$ successes and $n - x$ failures are exchangeable: for any length, the probability is proportional to $\theta^x(1-\theta)^{n-x}$
- Can you think of an exchangeable sequence that is not iid?
- Suppose I assume that if I toss a coin twice I can get 2H, 1H or 0H, and each is equally probable. This data generating process presumably corresponds to $P(HH) = 1/3, P(TT) = 1/3, P(TH) = P(HT) = 1/6$. These outcomes are exchangeable but not of the form $\theta^x(1-\theta)^{n-x}$

# Exchangeability

- DeFinetti's Exchangeability Theorem:
  $\exists \theta \geq 0$ such that $\int_0^1 p(\theta)d\theta = 1$ and

$$p(x, n-x) = \int_0^1 \theta^x (1-\theta)^{n-x} p(\theta) d\theta$$

- This $\theta$, whose existence is guaranteed for exchangeable sequences, can be interpreted as the Bayesian prior. But notice that it is not imposed into the problem! Its existence follows from a symmetry assumption weaker than $iid$.

# Sketch proof of exchangeability

- Let $p_{k,n}$ denote $P(X, n-x)$ where $X_1, ..., X_m$ is an exchangeable sequence of Bernoulli random variables.
- Let $q_r = P(\sum_{i=1}^{m} X_i = r)$
- Then,
$$p_{k,n} = \sum_{r=0}^{m} \frac{(r)_k (m-r)_{n-k}}{(m)_n}$$
where $(x)_k = \prod_{j=0}^{k-1}(x-j)$
- From exchangeability, it follows that given $r$ ones, the distribution of $X_1, ..., X_m$ is the same as that obtained by drawing from an urn containing $r$ ones and $m-r$ zeros.
- Thus, the $r$th term of the series is
$$P\left[X_1 = 1, ..., X_r = 1, X_{r+1} = 0, ..., X_m = 0 | \sum_{j=1}^{m} X_j = r\right] \times P\left[\sum_{j=1}^{m} X_j = r\right]$$
- So we can rewrite the first eq. as
$$p_{k,n} = \int_0^1 \frac{(\theta m)_k ((1-\theta)m)_{n-k}}{(m)_n} F_m(d\theta)$$
- where $F_m$ is the distribution function concentrated on $\{r/m : 0 \leq r \leq m\}$ whose jump at $r/m$ is $q_r$.

# From exchangeability to inference

- DeFinetti's exchangeability theorem is predated by W.E. Johnson in 1924, known as Johnson's Sufficientness postulate. As part of this, he identifies an exchangeability notion which he called the Permutation Postulate.
- So the theorem says that exchangeable sequences are mixtures of Bernoulli sequences; the mixture being by a distribution over the latent value $\theta$.
- Let us make a simple but common Bayesian assumption about the distribution of $\theta$. In particular, given exchangeable Bernoulli trials of the sort we are dealing with, we can suppose that,
- $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- This distribution can be normalized by adding in front

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

  where $\Gamma(x)$ is the complete Gamma function, $\int_0^\infty t^{x-1}e^{-t}dt$
- Now let's use Bayes Theorem to get an answer!

# Bayesian approach

- If we assume that $\theta \sim \text{Beta}(\alpha, \beta)$, and $X \sim \text{Binomial}(n, p)$, then

$$\begin{aligned} p(\theta|x) &\propto p(\theta) f(x|\theta) \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{x}(1-\theta)^{n-x} \\ &\propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\binom{n}{x}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \\ &= \frac{\Gamma((x+\alpha)+(n-x+\beta))}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \end{aligned}$$

- Thus: the posterior is $\text{Beta}(\alpha + x, \beta + n - x)$
- How to choose $\alpha$ and $\beta$?

# Conjugate families

- Definition: A prior distribution is a member of the natural conjugate family if it is proportional to a kernel function of the same form as the likelihood:

$$p(\theta) \propto f(x|\theta)$$

for $p(\theta) \geq 0$ and $\int p(\theta)d\theta = 1$

- The beta distribution is conjugate to the binomial

# Mean and variance

- Prior mean: $\alpha/(\alpha+\beta)$
- Posterior mean: $(\alpha+x)/(\alpha+\beta+n)$
- Prior variance:

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- Posterior variance:

$$\frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$

# Predictions

- Predictive distribution:

$$p(\widetilde{X} = 1|x_1, ..., x_n) = \int p(\widetilde{X} = 1|x_1, ..., x_n, \theta)p(\theta|x_1, ..., x_n)d\theta$$
$$= \mathrm{E}[\theta|x_1, ..., x_n]$$

- So the probability that the 13th coin toss lands on heads is

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

- $\alpha$, $\beta$ are pseudo trials
- When $\alpha = \beta = 1$, we have

$$\frac{x + 1}{n + 2}$$

- This is Laplace's rule of succession! It is the special case of the posterior predictive distribution, given a beta-binomial model, under a uniform prior for the coin's bias.

# Predictions

- If we use the posterior mean to estimate the coin's bias, is the estimate asymptotically consistent?

$$\lim_{n\to\infty}\left[\mathrm{E}[\theta|x]-\frac{x}{n}\right]=0$$

- Is it biased?
- Revisit: How to choose $\alpha$ and $\beta$?
- Note that the posterior mean is a compromise between prior information and data:

$$\mathrm{E}[\theta|x]=\frac{\alpha+x}{\alpha+\beta+n}=\frac{n}{\alpha+\beta+n}\left(\frac{y}{n}\right)+\frac{\alpha+\beta}{\alpha+\beta+n}\left(\frac{\alpha}{\alpha+\beta}\right)$$

- Rudolf Carnap called this the continuum of inductive method.
- On the Bayesian approach, no assumptions about stopping rule. However, some assumptions about symmetry (exchangeability).

# Example

- Estimate the chance $\theta$ of recidivism based on a study in which there were $n = 43$ individuals released from jail and $x = 15$ committed another crime within 36 months of release.
- Using a $\text{Beta}(2, 8)$ prior for $\theta$, find the posterior mean, standard deviation, and $95\%$ credible interval
- Note: $\int_a^b p(\theta|x)d\theta = x$ is the $(x \times 100)\%$ credible interval.
- $\theta|x = 15 \sim \text{Beta}(17, 36)$
- $\text{E}[\theta|x = 15] = 17/53 = 0.32$
- $\text{Sd}(\theta|x = 15) = [(17 \times 36)/(53^2 \times 54)]^{1/2} = 0.06$
- $\text{CI}(\theta|x = 15) = (0.20, 0.45)$
- R code for CI: qbeta(c(0.025,0.975),17,36)

# Normal processes

- Important in many statistical modeling problems
- Often useful as approximation or a component in more complicated models
- We will treat separately cases with known variance and known mean.

# Unknown mean, known variance

- Suppose $x_i|\mu \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$ with $\sigma^2$ known. Let $x = (x_1, ..., x_n)$.
- What is the likelihood?

$$p(x|\mu) = (2\pi\sigma^2)^{n/2} \exp\left[ -(2\sigma^2)^{-1} \sum_{i=1}^{n} (x_i - \mu)^2 \right]$$

- What is the natural conjugate prior?

# Unknown mean, known variance

- First, simplify the likelihood as follows:

$$p(x|\mu) \propto \exp\left[a(x)\mu^2 + b(x)\mu + c(x)\right]$$
$$\propto \exp\left[A\mu^2 + B\mu + C\right]$$

- Note that

$$A = -n(2\sigma^2)^{-1}$$

$$B = \sigma^{-2}\sum_{i=1}^{n} x_i$$

$$C = -(2\sigma^2)^{-1}\sum_{i=1}^{n} x_i^2$$

- To specify natural conjugate prior, set

$$p(\mu) \propto \exp\left[a^*\mu^2 + b^*\mu + c^*\right] = \exp\left[-\frac{1}{2\tau_0^2}(\mu - \mu_0)^2\right]$$

- This implies that $\mu \sim \mathrm{N}(\mu_0, \tau_0^2)$ with hyperparameters $\mu_0$ and $\tau_0^2$.

# Unknown mean, known variance



- The conjugate prior implies that the posterior for $\mu$ is exponential with a quadratic form – i.e., also normal.
- Note that in the posterior everything except $\mu$ is constant.

  So we write:

$$p(\mu|x) \propto \exp\left[-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\tau_0^2}\right)\right]$$

- Expanding the exponents, collecting terms, and completing the square in $\mu$, we get

$$p(\mu|x) \propto \exp\left[-\frac{1}{2\tau_1^2}(\mu-\mu_1)^2\right]$$

- Therefore, $\mu|x \sim \mathrm{N}(\mu_1, \tau_1^2)$ where

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}x}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \qquad \text{and} \qquad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

# Interpretation

- The posterior precision $\frac{1}{\tau_1^2}$ equals the prior precision plus the data precision.
- Several ways to interpret the posterior mean $\mu_1$:
  - The equation above suggests: a weighted average of the prior mean and the observed value, $x$, with weights proportional to the precisions.
  - Or, the prior mean adjusted toward $x$:

$$\mu_1 = \mu_0 + (x - \mu_0)\frac{\tau_0^2}{\sigma^2 + \tau_0^2}$$

  - In either case, and as with the beta-binomial model, the posterior is a "Carnapian" compromise between the prior mean and the observed value.

# Interpretation

- The point is, that we start from the prior and gradually update in response to data, where our responsiveness to the evidence is something that can be tuned through the hyperparameters.

# Normal predictions

- The posterior predictive distribution is,

$$p(\widetilde{x}|x) = \int p(\widetilde{x}|\mu)p(\mu|x)d\mu$$

- This is because the distribution of future observations does not depend on past data, *conditional* on $\mu$ (observations are conditionally independent given $\mu$)
- The right hand side is proportional to

$$\int \exp\left[-\frac{1}{2\sigma^2}(\widetilde{x}-\mu)^2\right]\exp\left[-\frac{1}{2\tau_1^2}(\mu-\mu_1)^2\right]d\mu$$

- To derive the product in the integrand, note that the joint density of $(\widetilde{x},\mu)^T$ must be bivariate normal (since we had a normal prior and likelihood), so the marginal density $\widetilde{x}|x$ must be normal.

# Normal predictions

- Posterior predictive mean:

$$\mathrm{E}[\widetilde{x}|x] = \mathrm{E}[\mathrm{E}[\widetilde{x}|\mu, x]|x] = \mathrm{E}[\mu|x] = \mu_1$$

- Posterior predictive variance:

$$\begin{aligned}\mathrm{Var}(\widetilde{x}|x) &= \mathrm{E}[\mathrm{var}(\widetilde{x}|\mu, x)|x] + \mathrm{Var}(\mathrm{E}[\widetilde{x}|\mu, x]|x) \\ &= \mathrm{E}[\sigma^2|x] + \mathrm{Var}(\mu|x) \\ &= \sigma^2 + \tau_1^2\end{aligned}$$

# Known mean, unknown variance

- Let $x_i \sim \mathrm{N}(\mu, \sigma^2)$ with $\mu$ known and $\sigma^2$ unknown
- The likelihood function is,

$$p(x|\sigma^2) = (2\pi\sigma^2)^{n/2} \exp\left[ -(2\sigma^2)^{-1} \sum_{i=1}^{n} (x_i - \mu)^2 \right]$$

- For a natural conjugate prior, need prior with same functional form,

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp\left[ -\frac{\beta}{\sigma^2} \right]$$

- This is the kernel of an inverse Gamma distribution with hyperparameters $\alpha$ and $\beta$. Thus $\sigma^2 \sim G^{-1}(\alpha, \beta)$
- The normalizing constant is $\frac{\beta^\alpha}{\Gamma(\alpha)}$.
- Posterior:

$$p(\sigma^2|x) \propto (\sigma^2)^{-n/2+\alpha+1} \exp\left[ -\frac{v/2 + \beta}{\sigma^2} \right]$$

  where $v = n^{-1} \sum_{i=1}^{n} (x_i - \mu)^2$
- Thus, $\sigma^2|x \sim \mathrm{G}^{-1}(n/2, v/2)$ and $\sigma^{-2}|x \sim \mathrm{G}(\alpha + n/2, \beta + v/2)$

# Example

A random sample of $n$ students is drawn from a large population, and their weights are measured. The average weight of the $10$ sampled students is 140 pounds. Assume the weights in the population are normally distributed with unknown mean $\mu$ and known standard deviation $20$ pounds. Suppose your prior distribution for $\mu$ is normal with mean $180$ and standard deviation $40$.

1. Find the posterior distribution for $\mu$
2. A new student is sampled at random from the same population and has a weight of $\widetilde{x}$ pounds. Give a posterior predictive distribution for $\widetilde{x}$.
3. Find a 95% posterior credible interval for $\mu$ and 95% posterior predictive interval for $\widetilde{x}$.
4. Suppose you obtain additional $20$ samples with a mean of $155$ pounds. Using all samples repeat (a), (b) and (c).

# Part A

- The sampling distribution for $x_i$ is

$$x_i \sim \mathrm{N}(\mu, 400),\ i = 1, ..., n$$

- The prior distribution for $\mu$ is

$$\mu \sim \mathrm{N}(180, 1600)$$

- The posterior distribution for $\mu$ is

$$\mu|x \sim \mathrm{N}\left(\frac{4n\overline{x} + 180}{4n + 1}, \frac{1600}{4n + 1}\right)$$

  where $x = (x_1, ..., x_n)$ and $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$

- When $n = 10$ and $\overline{x} = 140$, the posterior distribution is

$$\mu|x \sim \mathrm{N}(141.0, 39.0)$$

# Part B



- The posterior predictive distribution is

$$\widetilde{x}|x \sim \mathrm{N}\left(\frac{4n\overline{x}+180}{4n+1}, \frac{1600n+2000}{4n+1}\right)$$

- With $n = 10$ and $\overline{x} = 140$, the posterior predictive distribution is

$$\mu|x \sim \mathrm{N}(141.0, 439.0)$$

# Part C

- The 95% credible interval for $\mu$ given $n = 10, \overline{x} = 140$:

$$141.0 \pm 1.96(6.25) = [128.7, 153.2]$$

- The 95% credible interval for $\widetilde{x}$ given $n = 10, \overline{x} = 140$:

$$141.0 \pm 1.96(20.95) = [99.9, 182.0]$$

# Part D

- With additional $20$ samples with a mean of $155$ pounds, the total sample size $n = 30$ with a mean of $\overline{x} = (20(155) + 10(140))/30 = 150$.
- The posterior distribution for $\mu$ is

$$\mu|x \sim \mathrm{N}(150.2, 13.2)$$

- The posterior predictive distribution for $\widetilde{x}$ is

$$\widetilde{x}|x \sim \mathrm{N}(150.2, 413.2)$$

- The 95% credible interval for $\mu$ given $\overline{x} = 150, n = 30$ is

$$150.2 \pm 1.96(3.64) = [143.1, 157.4]$$

- The 95% credible interval for $\widetilde{x}$ given $\overline{x} = 150, n = 30$ is

$$150.2 \pm 1.96(20.32) = [110.4, 190.1]$$

# Poisson models

- Recall that $X \sim$ Poisson if

$$Pr(X = k|\theta) = \frac{\theta^k}{k!} \exp(-\theta),\ k \in \{0, 1, 2, ..., \}$$

- Suppose $(x_1, ..., x_n) \sim$ Poisson with mean $\theta$. Then the joint data pmf is

$$p(x|\theta) \propto \exp\left[a(x)\theta + b(x)\log(\theta)\right]$$

  where $a(x) = -n$ and $b(x) = \sum_{i=1}^{n} x_i$
- The natural conjugate prior given this form is Gamma.
- If $x_i \sim \text{Poisson}(\theta)$ and $\theta \sim \text{G}(\alpha, \beta)$, the posterior is

$$\theta|x \sim \text{G}(\alpha + n\overline{x}, \beta + n)$$

  where $x = (x_1, ..., x_n)$ and $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$

# Poisson models

- The posterior mean and variance are:

$$\mathrm{E}[\theta|x] = (\alpha + n\overline{x})(\beta + n)^{-1}$$

$$\mathrm{Var}(\theta|x) = (\alpha + n\overline{x})(\beta + n)^{-2}$$

- As before, posterior mean is a compromise:

$$\mathrm{E}[\theta|x] = \frac{\beta}{\beta + n}\frac{\alpha}{\beta} + \frac{n}{\beta + n}\overline{x}$$

- How to interpret the hyperparameters?
  - $\beta$: the number of prior observations
  - $\alpha$: sum of counts form $\beta$ prior observations