Bayesian methods in economics and finance

Common distributions and conjugate priors
Your project work

► should discuss small question in Bayesian statistics or illustrate possible application/case study
► must be closely related to topics covered in the lecture
► should contain own computation – analytic or numerical
► should be structured like a 5-10 page paper and must contain about 3-5 references
► is due by Aug 31

Please contact me to discuss possible topics
Binomial distribution

Recall coin tossing example:

- **Binomial distribution**:
  
  \[ p(k|n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \]

- **Sampling space**: \( \mathcal{K} = \{0, \ldots, n\} \)
- **Parameter space**: \( \Theta = [0, 1] \)
- **Mean**: \( n\theta \)
- **Variance**: \( n\theta(1 - \theta) \)

- **Conjugate prior is Beta distribution**:
  
  \[ p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1} \]
Poisson distribution

Used to model number of independent events occurring in unit time interval:

\[ p(k|\lambda) = e^{-\lambda} \frac{\lambda^k}{k!} \]

- Sample space: \( \mathcal{K} = \mathbb{N} \)
- Parameter space: \( \Lambda = \mathbb{R}_{>0} \)
  Positive rate parameter \( \lambda \)
- Mean: \( \lambda \)
- Variance: \( \lambda \)
- Arises as limit of binomial distribution as \( n \to \infty \) and \( \theta \to 0 \)
  such that \( \lambda = n \theta_n \) constant

**Q:** Can you find the conjugate prior?
Conjugate prior for rate parameter $\lambda$ of Poisson distribution:

$$p(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta \lambda}$$

- Sampling space: $\Lambda = (0, \infty)$
- Parameter space: $\alpha > 0$ (shape) and $\beta > 0$ (rate)
- Mean: $\frac{\alpha}{\beta}$
- Variance: $\frac{\alpha}{\beta^2}$
Normal distribution

\[ p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \]

- Sampling space: \( \mathcal{X} = \mathbb{R} \)
- Parameter space: \( \mu \in \mathbb{R} \) and \( \sigma \in \mathbb{R}_{>0} \)
- Mean: \( \mu \)
- Variance: \( \sigma^2 \)
Normal distribution

Often, it is more convenient to use a different parametrization:

\[ p(x|\mu, \tau) = \left( \frac{\tau}{2\pi} \right)^{\frac{1}{2}} e^{-\frac{1}{2} \tau (x-\mu)^2} \]

with precision \( \tau = \frac{1}{\sigma^2} \)

Next, we consider

- Estimating \( \mu \) when \( \tau \) is known
- Estimating \( \tau \) when \( \mu \) is known
- Estimating \( \mu \) and \( \tau \) together
Mean estimation

Conjugate prior for $\mu$ is again a normal distribution:

$$p(\mu|\mu_0, \tau_0) = \left(\frac{\tau_0}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2} \tau_0 (\mu - \mu_0)^2}$$

Posterior given data $D = (x_1, \ldots, x_N)$ is found by “completing the square”:

$$p(\mu|D) \propto \prod_{i=1}^{N} p(x_i|\mu, \tau)p(\mu|\mu_0, \tau_0)$$

$$\propto e^{-\frac{1}{2} \tau \sum_i (x_i - \mu)^2} e^{-\frac{1}{2} \tau_0 (\mu - \mu_0)^2}$$

$$= e^{-\frac{1}{2} (\tau \sum_i x_i^2 - 2\tau \mu \sum_i x_i + N\tau \mu^2 + \tau_0 \mu^2 - 2\tau_0 \mu \mu_0 + \tau_0 \mu_0^2)}$$

$$\propto e^{-\frac{1}{2} ((N\tau + \tau_0)\mu^2 - 2(\tau \sum_i x_i + \tau_0 \mu_0)\mu)}$$

$$\propto e^{-\frac{1}{2} (N\tau + \tau_0)(\mu - \frac{\tau \sum_i x_i + \tau_0 \mu_0}{N\tau + \tau_0})^2}$$
Mean estimation

Conjugate prior for $\mu$ is again a normal distribution:

$$p(\mu|\mu_0, \tau_0) = \left(\frac{\tau_0}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2} \tau_0 (\mu - \mu_0)^2}$$

Posterior given data $D = (x_1, \ldots, x_N)$ is found by “completing the square”:

$$p(\mu|D) \propto \prod_{i=1}^{N} p(x_i|\mu, \tau) p(\mu|\mu_0, \tau_0)$$

$$\propto e^{-\frac{1}{2} (N\tau + \tau_0)(\mu - \frac{\tau \sum_i x_i + \tau_0 \mu_0}{N\tau + \tau_0})^2}$$

Normal distribution with

- precision $\tau_D = N\tau + \tau_0$
- mean

$$\mu_D = \frac{\tau \sum_i x_i + \tau_0 \mu_0}{N\tau + \tau_0} = \frac{1}{\tau_D} (N\tau \hat{\mu} + \tau_0 \mu_0)$$

where $\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$
Precision estimation

Now, assume that the mean $\mu$ is known and we want to infer the precision $\tau$:

$$p(\tau|D, \mu) \propto p(\tau) \prod_{i=1}^{N} p(x_i|D, \mu, \tau)$$

$$= p(\tau) \left( \frac{\tau}{2\pi} \right)^{\frac{N}{2}} e^{-\frac{1}{2} \tau \sum_i (x_i - \mu)^2}$$

Thus, $\tau$ occurs in the likelihood as $\tau^{\frac{N}{2}} e^{-\frac{1}{2} \sum_i (x_i - \mu)^2 \tau}$ and the conjugate prior is a gamma distribution:

$$p(\tau|D, \mu) \propto \tau^{\alpha - 1} e^{-\beta \tau} \tau^{\frac{N}{2}} e^{-\frac{1}{2} \sum_i (x_i - \mu)^2 \tau}$$

$$= \tau^{\alpha' - 1} e^{-\beta' \tau}$$

where $\alpha' = \alpha + \frac{N}{2}$ and $\beta' = \beta + \frac{1}{2} \sum_i (x_i - \mu)^2$ are the posterior Gamma parameters.
Student’s t-distribution

\[ p(x|\mu, \sigma^2, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\sigma^2\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{1}{\nu} \frac{(x - \mu)^2}{\sigma^2}\right)^{-\frac{\nu+1}{2}} \]

- Sampling space: \( \mathcal{X} = \mathbb{R} \)
- Parameter space: \( \mu \in \mathbb{R} \) (mean), \( \sigma^2 > 0 \) (variance) and \( \nu > 0 \) (degrees of freedom)
- Mean: \( \mu \) for \( \nu > 1 \)
- Variance: \( \sigma^2 \frac{\nu}{\nu - 2} \) for \( \nu > 2 \)

Marginal distribution of Gaussian with unknown precision:

\[ p(x|\mu, \alpha, \beta) = \int_0^{\infty} \mathcal{N}(x|\mu, \tau)\text{Gamma}(\tau|\alpha, \beta) \, d\tau \]

\[ = \text{Stud}(x|\mu, \frac{\beta}{\alpha}, 2\alpha) \]
Student’s t-Distribution

Robustness to outliers: Gaussian vs t-distribution.
Joint estimation

To compute the joint posterior \( p(\mu, \tau | D) \) we proceed similarly and investigate the form of the likelihood:

\[
p(D|\mu, \tau) \propto \tau^{\frac{N}{2}} e^{-\frac{1}{2} \sum_i (x_i - \mu)^2 \tau}
\]

\[
= \tau^{\frac{N}{2}} e^{-\frac{1}{2} \tau (\sum_i x_i^2 - 2\mu \sum_i x_i + N\mu^2)}
\]

\[
= \tau^{\frac{N}{2}} e^{-\frac{1}{2} (\sum_i x_i^2 - N\hat{\mu}^2) \tau} e^{-\frac{1}{2} N\tau (\mu - \hat{\mu})^2}
\]

\[
= \tau^{\frac{N}{2}} e^{-\frac{1}{2} (\sum_i (x_i - \hat{\mu})^2) \tau} e^{-\frac{1}{2} N\tau (\mu - \hat{\mu})^2}
\]

where \( \hat{\mu} = \frac{1}{N} \sum_i x_i \)

Considering the above form as the product \( p(\tau)p(\mu | \tau) \) we can identify a suitable prior:

\[
p(\tau)p(\mu | \tau) \propto \tau^{\frac{n}{2}} e^{-\beta \tau} e^{-\frac{1}{2} \eta \tau (\mu - \mu_0)^2},
\]

i.e. the product of a Gamma prior \( p(\tau | \alpha, \beta) \) and a normal prior \( p(\mu | \mu_0, \tau') \) where \( \alpha = \frac{n}{2} + 1 \) and \( \tau' = \eta \tau \).
Joint estimation

Using R, we can illustrate the joint inference as follows:

```r
normal.gamma <- function (mu, tau, mu0, eta, beta) {
  dgamma(tau, shape=eta/2 + 1, rate=beta)*dnorm(mu, mean=mu0, sd=1/sqrt(eta*tau))
}
N <- 10
x <- rnorm(N, mean=0.5, sd=0.25)
mu.hat <- mean(x)
alpha.0 <- 1; beta.0 <- 1; mu.0 <- 0
eta.0 <- 2*(alpha.0-1); eta.D <- eta.0 + N;
mu.D <- 1/eta.D*(eta.0*mu.0 + N*mu.hat)
beta.D <- beta.0 + 0.5*(sum((x-mu.hat)^2) + eta.0/eta.D*(mu.hat-mu.0)^2)
mu <- seq(0,1,0.01); tau <- seq(1,25,0.1)
contour(mu, tau, outer(mu, tau, FUN=function(mu,tau) { normal.gamma(mu,tau,mu.D,eta.D,beta.D) })))
points(mean(x), 1/mean((x-mean(x))^2))
```

- Mean and standard deviation are not independent
- Normal-Gamma distribution often specified with separate parameters \( \eta \) and \( \alpha \)
Multi-variate normal

Generalizes normal distribution to $d$ dimensions:

$$p(x|\mu, \Sigma) = (2\pi)^{-\frac{d}{2}} \det |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)}$$

- Sampling space: $x \in \mathbb{R}^d$
- Parameter space: $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$
- Mean: $\mathbb{E}[x] = \mu$, i.e. $\mathbb{E}[x_i] = \mu_i$
- Covariance matrix: $\mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])] = \Sigma_{ij}$
  Matrix notation: $\Sigma = \mathbb{E}[(x - \mu)(x - \mu)^T]$
Multi-variate normal

Properties of the multi-variate normal distribution:

- Precision matrix: $\Lambda = \Sigma^{-1}$
- With diagonal covariance matrix, i.e. $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$:

$$p(x|\mu, \Sigma) = \prod_{i=1}^{d} \left(2\pi\sigma_i^2\right)^{-\frac{1}{2}} e^{-\frac{1}{2} \frac{(x_i-\mu_i)^2}{\sigma_i^2}},$$

i.e. reduces to a product of $d$ 1-dimensional Gaussians

- Iso-probability lines are ellipses defined by the quadratic form

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = \text{const}$$

Most easily seen in 2-dimensions (assuming $\mu = 0$):

- Diagonal covariance:

$$(x_1 x_2) \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} = \text{const}$$

Recall: $x_1^2 + x_2^2 = 1$ defines unit circle

- General covariance matrix rotates axis. Elongation of ellipses is longest when precision is low
Multi-variate normal

Properties of the multi-variate normal distribution:

- Precision matrix: $\Lambda = \Sigma^{-1}$
- With diagonal covariance matrix, i.e. $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$:

$$p(x|\mu, \Sigma) = \prod_{i=1}^{d} (2\pi \sigma_i^2)^{-\frac{1}{2}} e^{-\frac{1}{2} \frac{(x_i-\mu_i)^2}{\sigma_i^2}},$$

i.e. reduces to a product of $d$ 1-dimensional Gaussians

- Iso-probability lines are ellipses defined by the quadratic form

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = \text{const}$$

- Consider $x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}$. Then,

  - Marginal distribution $p(x_a)$ is again Gaussian with mean $\mu_a$ and covariance matrix $\Sigma_{aa}$
  - Conditional distribution $p(x_a|x_b)$ is again Gaussian with mean $\mu_a|b = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab} (x_b - \mu_b)$ and precision $\Lambda_{a|b} = \Lambda_{aa}$
Show that the conditional distribution \( p(x_a|x_b) \) is Gaussian:

\[
p(x_a|x_b) \propto p(x_a, x_b)
\]

\[
\propto \exp \left\{ -\frac{1}{2} \begin{pmatrix} x_a - \mu_a \\ x_b - \mu_b \end{pmatrix}^T \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \begin{pmatrix} x_a - \mu_a \\ x_b - \mu_b \end{pmatrix} \right\}
\]

\[
\propto \exp \left\{ -\frac{1}{2} (x_a^T \Lambda_{aa} x_a - 2x_a^T (\Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b))) \right\}
\]

\[
\propto \exp \left\{ -\frac{1}{2} (x_a - \mu_{a|b})^T \Lambda_{aa} (x_a - \mu_{a|b}) \right\}
\]
Wishart distribution

\[
p(\mathbf{X}|\mathbf{V}, n) = \frac{1}{2^{np/2} \det|\mathbf{V}|^{n/2} \Gamma_p(n/2)} \det|\mathbf{X}|^{n-p-1/2} e^{-\frac{1}{2} \text{tr}(\mathbf{V}^{-1}\mathbf{X})}
\]

where \( \Gamma_p(n/2) = \pi^{p(p-1)/4} \prod_{j=0}^{p-1} \Gamma\left(\frac{n-j}{2}\right) \)

- Sampling space: \( \mathbf{X} \in \mathbb{R}^{p \times p} \)
- Parameter space: \( \mathbf{V} \in \mathbb{R}^{p \times p} \) and degrees of freedom \( n > p + 1 \)
- Mean: \( n\mathbf{V} \)
- Variance: \( n(V_{ij}^2 + V_{ii} V_{jj}) \)
Wishart distribution

- Conjugate prior for the precision $\Lambda$ of multi-variate Gaussian with known mean $\mu$

- Sampling distribution of empirical covariance matrix, i.e. $\hat{\Sigma} = \sum_{i=1}^{N} xx^T \sim \text{Wishart}(\Sigma, n)$ when $x_i \sim \text{Normal}(0, \Sigma)$

- Posterior $p(\Lambda|D)$ is Wishart distribution with

$$\Lambda_D = (\Lambda_0^{-1} + \hat{\Sigma})^{-1} \quad \text{and} \quad n_D = n + n_0$$

- Unknown mean and precision:

$$p(\mu, \Lambda) = \text{Normal}(\mu|\mu_0, (\eta\Lambda)^{-1})\text{Wishart}(\Lambda|\Lambda_0, n_0)$$

*Normal-Wishart distribution* as conjugate prior
Choosing priors

How should we choose a suitable prior?

- Mathematical convenience: Conjugate prior
- Let the data speak: Uninformative priors
- Invariance principle: Jeffreys prior
Conjugate priors

Conjugate priors are often a good choice as
- Posterior is analytically tractable
- Parameters often interpretable as pseudo-observations

But, can correspond to strong assumptions that are hard to justify
Uninformative priors

What to do if we have little information a-priori?

▶ Natural guess: Uniform prior

But: Choice depends on parametrization:

▶ Consider parameter $\theta$ with prior $p(\theta)$
▶ Reparametrize as $\eta = h(\theta)$:

\[
p(\eta) = p(\theta)\left|\frac{d\theta}{d\eta}\right|
\]

\[
= p(h^{-1}(\eta))\left|\frac{d}{d\eta}h^{-1}(\eta)\right|
\]

Example: $\sigma = e^\theta$ with $p(\theta) \propto 1$:

\[
p(\sigma) \propto 1\left|\frac{d}{d\sigma}\log \sigma\right| = \frac{1}{\sigma}
\]
Uninformative priors

What to do if we have little information a-priori?

- Natural guess: Uniform prior
  - **But**: Choice depends on parametrization:
  - Better: Invariance principle
    - Prior for location parameter $\mu$
      
      $$p(x|\mu) = f(x - \mu) = p(x + c|\mu + c) \quad \forall c$$

      should be translation invariant, i.e.

      $$\int_a^b p(\mu)d\mu = \int_{a+c}^{b+c} p(\mu)d\mu = \int_a^b p(\mu + c)d\mu \quad \forall a, b, c$$

      Thus, $p(\mu) \propto 1$ uniform
Uninformative priors

What to do if we have little information a-priori?

- Natural guess: Uniform prior
  - **But**: Choice depends on parametrization:
  - Better: Invariance principle
    - Prior for *scale parameter* $\sigma$

$$p(x|\sigma) = \frac{1}{\sigma}f\left(\frac{x}{\sigma}\right)$$

should be *scale invariant*, i.e.

$$\int_a^b p(\sigma)d\sigma = \int_{ca}^{cb} p(\sigma)d\sigma = \int_a^b \frac{1}{c}p\left(\frac{\sigma}{c}\right)d\sigma \quad \forall a, b, c$$

Thus, $p(\sigma) \propto \frac{1}{\sigma}$

Note: Previous slide shows that $p(\log \sigma)$ uniform!
Improper priors

Prior $p(\theta)$ is called *improper* if it cannot be normalized, i.e.

$$\int_{\mathbb{R}} p(\theta) d\theta = \infty$$

- Uninformative priors $p(\mu) \propto 1$, $p(\sigma) \propto \frac{1}{\sigma}$ are improper
- Bayesian inference is still valid if posterior can be normalized:
  - Estimate mean of Gaussian sample $p(\mu|D)$
  - Conjugate prior $p(\mu|\mu_0, \tau_0)$ uniform for $\tau_0 \to 0$
  - Posterior $p(\mu|\mu_D, \tau_D)$ well defined in this limit:

$$\tau_D = \mathcal{N}\tau \text{ and } \mu_D = \hat{\mu}$$

Note: Posterior depends on data only
Jeffreys prior

Idea: Prior density should be unchanged under reparametrization:

\[ p(\theta) \propto \sqrt{I(\theta)} \]

where \( I(\theta) = \mathbb{E}[(\frac{d}{d\theta} \log p(x|\theta))^2] \) denotes the Fisher information

- Consider \( \eta = h(\theta) \)

\[
p(\eta) = p(\theta) \left| \frac{d\theta}{d\eta} \right|
\]

\[
\propto \sqrt{\mathbb{E}[(\frac{d}{d\theta} \log p(x|\theta))^2](\frac{d\theta}{d\eta})^2}
\]

\[
= \sqrt{\mathbb{E}[(\frac{d}{d\theta} \log p(x|\theta) \frac{d\theta}{d\eta})^2]}
\]

\[
= \sqrt{I(\eta)}
\]
Jeffreys prior

Examples:

- Gaussian distribution $p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}$:

  $$p(\mu) \propto \sqrt{\mathbb{E}[\left( \frac{d}{d\mu} \log p(x|\mu) \right)^2]}$$
  $$= \sqrt{\mathbb{E}[\left( \frac{x - \mu}{\sigma^2} \right)^2]}$$
  $$= \sqrt{\frac{\sigma^2}{\sigma^4}} \propto 1$$

- Binomial distribution $p(k|\theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$.
  Jeffreys prior is Beta distribution with $\alpha = \beta = \frac{1}{2}$.
Some remarks

- Uninformative priors are often employed
  - Often limits of conjugate priors with analytic posterior
  - Inference resembles frequentist estimates
- Not a good idea in high-dimensions
  - Where is the probability mass of a standard Gaussian in $\mathbb{R}^d$?
    - Thin shell around origin at radius $\sqrt{d}$.
    - Thus, uninformative prior puts infinite mass on infinity!
  - Inference in high-dimensions profits from informed priors:
    - Recall James-Stein phenomenon
    - Power of shrinkage and penalized maximum likelihood
Exchangeability

More on priors:

- Can existence be motivated from first principles?
- Structural assumptions expressed by priors?
Exchangeability

More on priors:
  ▶ Can existence be motivated from first principles?
  ▶ Structural assumptions expressed by priors?

**Exchangeability:**
  ▶ Expresses symmetries of probabilistic models
  ▶ Derives representation in terms of latent variables
Exchangeability

Consider an (infinite) sequence $X_1, X_2, \ldots$ of random variables.

**Definition**

A sequence $X_1, X_2, \ldots$ of random variables is called (infinitely) exchangeable if

$$(X_1, X_2, \ldots, X_n) \overset{d}{=} (X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)})$$

for all $n \in \mathbb{N}$ and all permutations $\pi$ of $1, \ldots, n$.

**Intuition:** Order of the sequence does not matter . . .

- Sequences of independent and identically distributed random variables (i.i.d.) are exchangeable,
- but **not** every exchangeable sequence is i.i.d.
Pólya urn

Consider an urn containing \( r \) red and \( b \) blue balls. Now repeat the following process:

▶ Draw a ball at random and note its color
▶ Place it back *together with an additional ball* of the same color

Obviously, consecutive draws are **not** independent, but the process is exchangeable:

\[
\begin{align*}
p(b, b, r) &= \frac{b}{r + b} \frac{b + 1}{r + b + 1} \frac{r}{r + b + 2} \\
p(b, r, b) &= \frac{b}{r + b} \frac{r}{r + b + 1} \frac{b + 1}{r + b + 2}
\end{align*}
\]
De Finetti theorem

De Finetti representation theorem:

**Theorem**

A binary sequence $X_1, X_2, \ldots$ is exchangeable if and only if there exists a measure $\mu$ on $[0, 1]$ such that for all $n$

$$p(x_1, \ldots, x_n) = \int_{\theta} \theta^{t_n} (1 - \theta)^{n-t_n} d\mu(\theta)$$

where $t_n = \sum_{i=1}^{n} x_i$.

Further,

- Given $\theta$ the sequence is i.i.d. Bernoulli distributed, i.e.

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1-x_i}$$

Think of $\theta$ as the bias of a coin.
De Finetti theorem

De Finetti representation theorem:

Theorem

A binary sequence $X_1, X_2, \ldots$ is exchangeable if and only if there exists a measure $\mu$ on $[0, 1]$ such that for all $n$

$$p(x_1, \ldots, x_n) = \int_\theta \theta^{t_n} (1 - \theta)^{n-t_n} d\mu(\theta)$$

where $t_n = \sum_{i=1}^{n} x_i$.

Further,

- Given $\theta$ the sequence is i.i.d. Bernoulli distributed.
- A law of large numbers holds, i.e.

$$\lim_{n \to \infty} \sum_{i=1}^{n} \frac{X_i}{n} \sim \mu$$
De Finetti theorem

De Finetti’s representation theorem can be interpreted in several ways:

- **Frequentist**: There is an unknown $\theta$ such that $X_1, X_2, \ldots$ are i.i.d. Bernoulli$(\theta)$ distributed. $\theta$ is the limiting frequency of observed 1’s (heads).

- **Bayesian**: Exchangeable distribution $P$ expresses beliefs/assumptions about $X_1, X_2, \ldots$. Observed distribution is permutation invariant and $\mu(\theta)$ is the subjective prior about the coin bias $\theta$.

- **Preferred**: Model observations $X_1, X_2, \ldots$ as a-priori alike
  - Implies a decomposition into structure $\theta$ (coin bias) and randomness $p(X_i|\theta)$ (coin flip)
  - Structural model is an (implicit) assumption $p(\theta)$. 

Exchangeability