Bayesian Inference and Decision Theory

Unit 2: Random Variables, Parametric Models, and Inference from Observation

GEORGE MASON UNIVERSITY

Learning Objectives for Unit 2

- Define and apply the following:
  - Event
  - Random variable (univariate and multivariate)
  - Joint, conditional and marginal distributions
  - Independence; conditional independence
  - Mass and density functions; cumulative distribution functions
  - Measures of central tendency and spread
- Use a graph to show conditional dependence and independence of random variables
- Be familiar with common parametric statistical models
  - Continuous distributions:
    - Normal, Gamma, Exponential, Uniform, Beta, Dirichlet
  - Discrete distributions:
    - Bernoulli, Binomial, Multinomial, Poisson, Negative Binomial
- Use Bayes rule to find the posterior and predictive distribution (marginal likelihood) of a parameter given a set of observations (using discrete approximation)
- State de Finetti's Theorem and its relationship to the concept of "true" probabilities

Random Variable: Definition

- Definition: A **random variable** is a function from the sample space Ω to a set of **outcomes** (often the real numbers)
  - Discrete (values can be enumerated) and continuous (values cannot be enumerated)
- Example:
  - Sample space Ω is a set of patients
  - Discrete random variable example: $Y(\omega)$ is 0 if patient ω has a fever (temperature > 37°C or 98.6°F) and 1 if patient doesn't have fever
  - Continuous random variable example: $X(\omega)$ is temperature of patient ω
- A random variable defines events on the outcome space
  - Example: Event Y = 1 is subset $\{\omega \in \Omega : Y(\omega) = 1\}$ consisting of patients with fever

RV from Frequentist View: Example

- Question: does "randomly chosen" patient have disease?
- Sample space S = {s1, s2, …} represents population of patients arriving at medical clinic
  - 30% of the patients in this population have the disease
  - 95% of the patients who have the disease test positive
  - 85% of the patients who do not have the disease test negative
- Random variables:
  - X maps randomly drawn patient s to X(s) = 0 if patient does not have disease and X(s) = 1 if patient has disease
  - Y maps randomly drawn patient s to Y(s) = 0 if patient's test result is negative and Y(s) = 1 if patient's test result is positive
- Usually we suppress the argument and write (X, Y) for the two-component random variable representing the condition and test result of a randomly drawn patient
  - Pr(X=1) = 0.3
  - Pr(Y=1 | X=1) = 0.95
  - Pr(Y=0 | X=0) = 0.85
  - Pr(X=1 | Y=1) = 0.73 (by Bayes rule)
- These probabilities refer to a *sampling process* and not to any particular patient

RV from Subjectivist View: Example

- Question: does a specific patient (e.g., George Smith) have disease?
- Sample space S = {s1, s2, …} represents "possible worlds" (or ways the world could be)
  - In 30% of these possible worlds, George has the disease
  - In 95% of these possible worlds in which George has the disease, the test is positive
  - In 85% of the possible worlds in which George does not have the disease, the test is negative
- Random variables:
  - X maps possible world s to X(s) = 0 if George does not have disease and X(s) = 1 if George has disease in this world
  - Y maps possible world s to Y(s) = 0 if George's test result is negative and Y(s) = 1 if George's test result is positive in this world
- Usually we suppress the argument and write (X, Y) for the two-component random variable representing George's condition and test result
  - Pr(X=1) = 0.3
  - Pr(Y=1 | X=1) = 0.95
  - Pr(Y=0 | X=0) = 0.85
  - Pr(X=1 | Y=1) = 0.73 (by Bayes rule)
- These probabilities refer to our *beliefs* about George's specific case

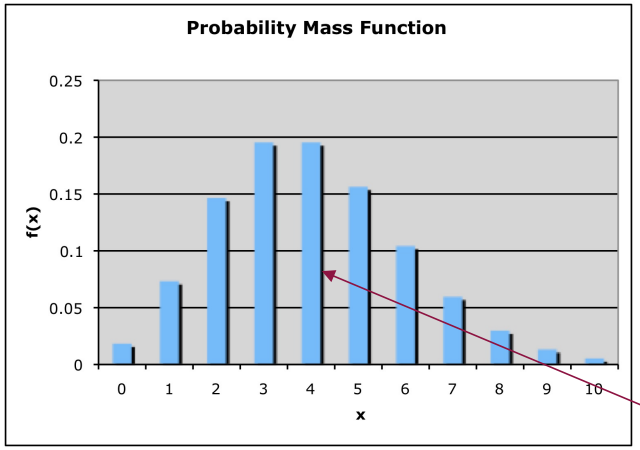
Comparison

- Frequentists say a probability distribution arises from some random process on the sample space (such as random selection)
  - Pr(X=1) = 0.3 means: "If patients are chosen at random from the population of clinic patients, 30% of them will have the disease"
  - Pr(X=1 | Y=1) = 0.73 means "If patients are chosen at random from the population of clinic patients who test positive, 73% of them will have the disease"
  - The probability George Smith has the disease is either 1 or 0, depending on whether he does or does not have the disease
- Subjectivists say a probability distribution represents someone's beliefs about the world
  - Pr(X=1) = 0.3 means: "If we have no information except that George walked into the clinic, then our degree of belief is 30% that he has the disease"
  - Pr(X=1 | Y=1) = 0.73 means "If we learn that George tested positive, our belief increases to 73% that he has the disease"
  - The probability that George has the disease depends on our information about George. If we know only that he went to the clinic, our probability is 30%. If we also learn that he tested positive, our probability becomes 73%. If we receive a definitive diagnosis, then the probability he has the disease is either 1 or 0, depending on whether he does or does not have the disease.

Mass, Density and Cumulative Distribution Functions for Random Variables

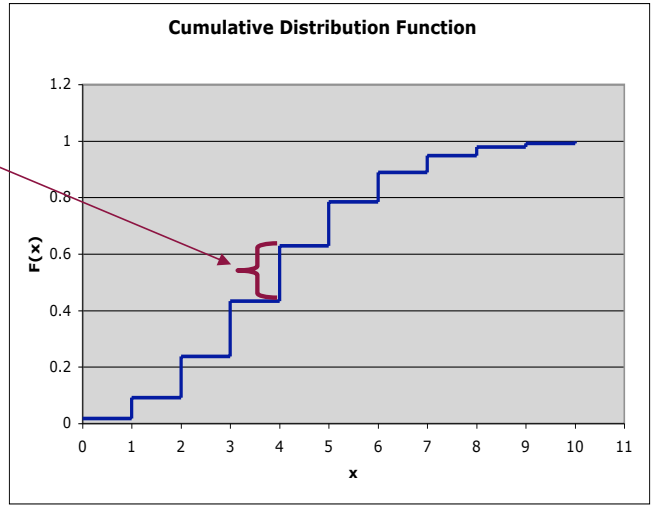
- **Cumulative distribution function** (cdf) - Value of cdf at *x* is probability that the random variable is less than or equal to *x*
- **Probability mass function** (pmf) - Maps each possible value of a discrete random variable to its probability
- **Probability density function** (pdf) - Maps each possible value of a continuous random variable to a positive number representing its likelihood relative to other possible values

Discrete Random Variable: pmf and cdf



Probability mass function

- f(x) is probability that random variable X is equal to x
- $\sum_x f(x) = 1$

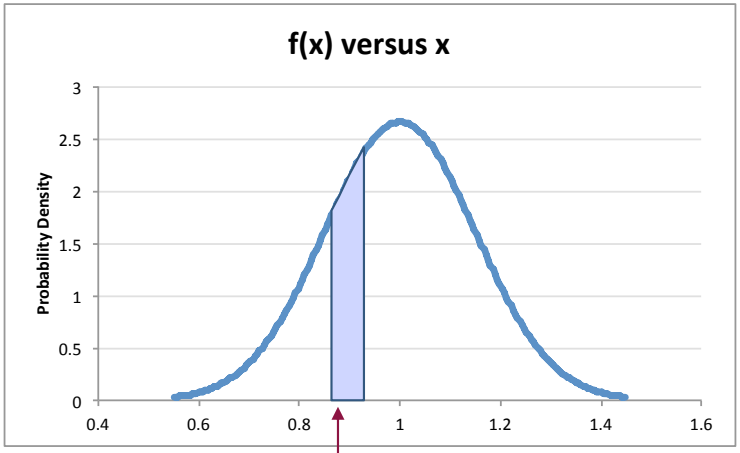


$$F(x) = \sum_{x' \le x} f(x')$$

Cumulative distribution function

- $F(x)$ is probability that random variable X is less than or equal to x
- Increases from zero (at -∞) to 1 (at +∞)
- Step function has jump at each value
- Height of step at possible value x is equal to value of pmf at x

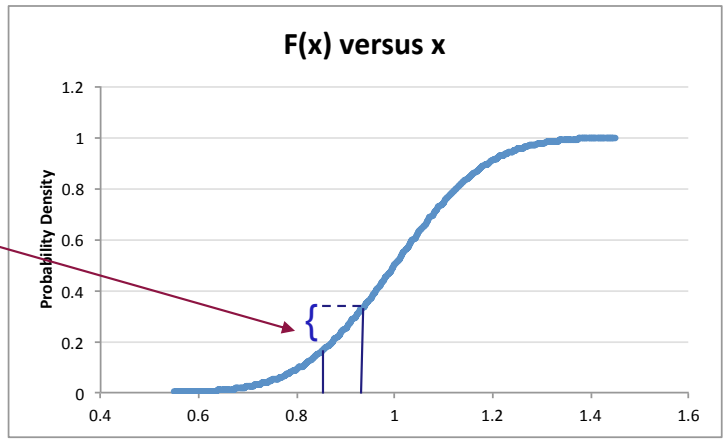
Continuous Random Variables: pdf and cdf



Area under f(x) between a and b is equal to difference. F(b)-F(a)

Probability density function

- Each individual value has probability zero
- pdf is derivative of cdf: $f(x) = dF(x)/dx$
- $f(x)\Delta x$ is approximately equal to the probability that X lies in $[x-\Delta x/2, x+\Delta x/2]$
- $\int_x f(x)dx = 1$



$$F(x) = \int_{x' \le x} f(x')dx'$$

Cumulative distribution function

- Probability that *X* lies in interval [a,b] is difference of cdf values and integral of pdf:

$$P(a < X \le b) = F(b) - F(a) = \int_a^b f(x)dx$$

Generalized Probability Density Function (gpdf)

- It is useful to have common notation for density and mass functions
- Integral symbol means integration when distribution is continuous and summation when distribution is discrete
  - Remember that an integral is a limit of sums
- gpdf f(x) is either a mass or density function:

$$\int_x g(x)f(x)d\mu(x) \equiv \sum_x g(x)f(x) \quad \text{if X is discrete}$$

$$\int_x g(x)f(x)d\mu(x) \equiv \int_x g(x)f(x)dx \quad \text{if X is continuous}$$

$\mu(x)$ stands for *measure* (*counting measure* for discrete random variables and *uniform measure* for continuous random variables)

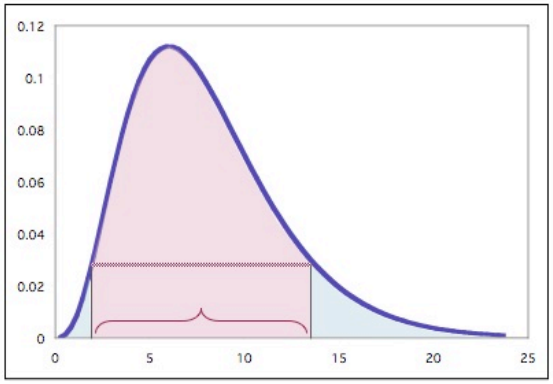
Central Tendency

- *Mean* (expected value E[*X*]): $E[X] = \int_x x f(x) d\mu(x)$
  - Weight each observation by its probability of occurring
  - In skewed distributions the mean is "pulled toward" the long tail
  - Sample mean of a random sample converges to the distribution mean as the sample size becomes large
- Median (50th percentile $x_{0.5}$): $x_{0.5} = F^{-1}(0.5) = \min_x\{F(x) \geq 0.5\}$
  - Half of the observations are larger, half smaller than the median
  - Less influenced by skewness and outliers than the mean
  - (There are several definitions of median of a discrete distribution)
- Mode (x_{max}): $f(x_{max}) \geq f(x)$ for all x
  - Most probable value
  - Distributions may have multiple modes (multiple peaks in pmf or pdf)
  - In multimodal distributions a single numerical summary of center may not be appropriate

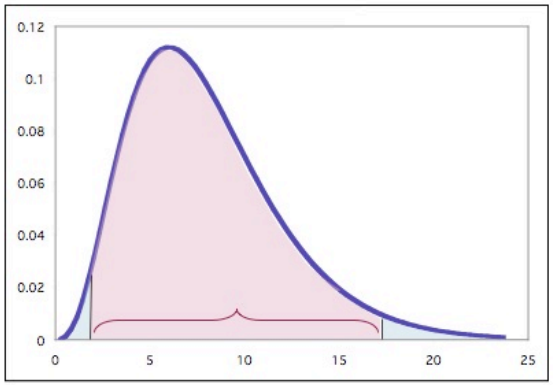
Spread

- Variance and standard deviation $V[X] = \int_x \left(x - E[X]\right)^2 f(x) d\mu(x)$
  - Variance is the expected value of the squared difference between an observation and the mean $SD[X] = \sqrt{V[X]}$
  - Standard deviation is the square root of the variance
  - Standard deviation is in same units as the variable itself
- Median absolute deviation $MAD[X] = |X - x_{0.5}|_{0.5}$
  - Median of the absolute value of the difference between an observation and the median
  - Less influenced by outliers than the standard deviation
- Credible interval (Bayesian confidence interval) $P\left(r \leq X \leq s\right) = \alpha$
  - Level-α credible interval [r,s]:
    - probability that the RV lies between r and s is α

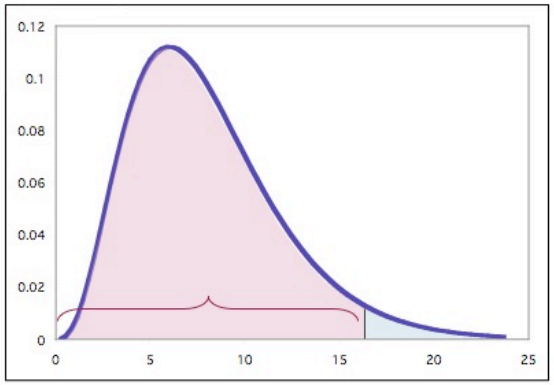
Varieties of Credible Interval



Highest density



Two-sided symmetric tail areas



One-sided

Theoretical and Sample Summary Statistics

- Expected value and sample mean

$$\mathrm{E}[X] = \int_x x f(x) d\mu(x) \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Theoretical and sample variance

$$\mathrm{Var}[X] = \int_x \left(x - \mathrm{E}[X]\right)^2 f(x) d\mu(x) \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

- Theoretical and sample quantiles / percentiles

$$F^{-1}(q) = \min_x \left\{ F(x) \geq q \right\}$$

  - There are several definitions for sample quantiles and quantiles of discrete distribution; this is one of the most common
  - What is the name of the 0.5 quantile?

Multivariate Random Variables

- Often we are uncertain about several random variables
- If they are related we can't treat them in isolation
- Examples:
  - Distribution of test result depends on whether patient has disease
  - Distribution of test score depends on socioeconomic status of student
- A *joint distribution* (also called a *multivariate distribution*) defines a probability distribution on several random variables

Formal Definition: Joint Distribution

- A *random vector* (or *multivariate random variable*) $\underline{X}$ is a vector of random variables
  - Underscore (often omitted) denotes a vector
- A *joint* distribution function is a function of several variables
  - cdf $F(\underline{x}) = P(\underline{X} \leq \underline{x})$
  - gpdf $f(\underline{x})$ (density if continuous, mass function if discrete)
  - Expected value vector

$$E[\underline{X}] = \int_{R^n} \underline{x} f(\underline{x}) d\mu(\underline{x})$$

  - Covariance matrix

$$Cov[\underline{X}] = \int_{R^n} (\underline{X} - E[\underline{X}])(\underline{X} - E[\underline{X}])^T f(\underline{x}) d\mu(\underline{x})$$

Marginal and Conditional Distributions

- If ($\underline{X}$,$\underline{Y}$) has joint gpdf f($\underline{x}$,$\underline{y}$):
  - The *marginal* gpdf of $\underline{X}$ is $f(\underline{x}) = \int_{\underline{y}} f(\underline{x},\underline{y})\, d\mu(\underline{y})$ ($\underline{y}$ is *marginalized out*)
  - The *conditional* gpdf of $\underline{X}$ given $\underline{Y}$ is $f(\underline{x} \mid \underline{y}) = \dfrac{f(\underline{x},\underline{y})}{f(\underline{y})}$
- The joint gpdf of $\underline{X}$ and $\underline{Y}$ can be factored into conditional and marginal gpdfs:

$$f(\underline{x},\underline{y}) = f(\underline{x} \mid \underline{y}) f(\underline{y}) = f(\underline{y} \mid \underline{x}) f(\underline{x})$$

- The joint gpdf for n random variables (X_1, …, X_n) can be written as:

$$f(x_1,\ldots x_n) = f(x_1) f(x_2 \mid x_1) f(x_3 \mid x_1, x_2) \cdots f(x_n \mid x_1,\ldots x_{n-1})$$

$$= \prod_{i=1}^{n} f(x_i \mid x_1,\ldots x_{i-1})$$

- For ***independent*** random variables the joint gpdf is the product of the individual gpdfs:

$$f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n)$$

Example: Joint, Conditional, Marginal Distributions

- Random variables:
  - X – possible values 0 (well) and 1 (sick)
  - Y – possible values 0 (tests negative) and 1 (tests positive)
- Joint distribution P(X,Y):

	X=0	X=1
Y=0	0.7 x 0.85 = 0.595	0.3 x 0.05 = 0.015
Y=1	0.7 x 0.15 = 0.105	0.3 x 0.95 = 0.285

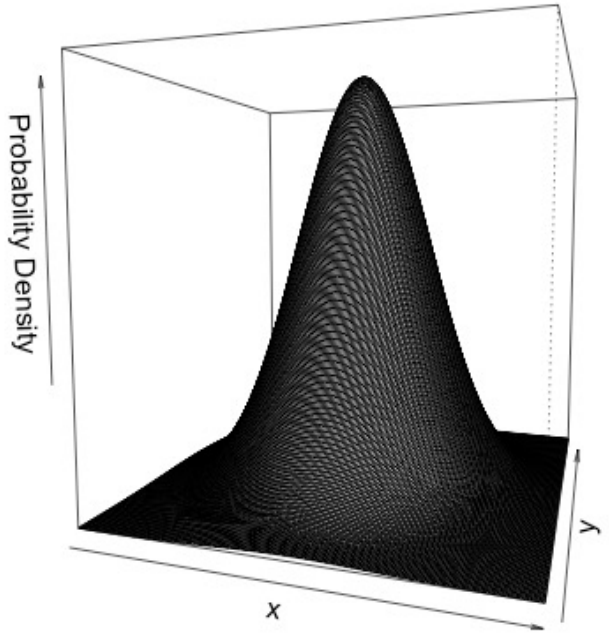
- Conditional distribution of Y given X:
  - P(Y=0 | X=0) = 0.85, P(Y=1 | X=0) = 0.15
  - P(Y=0 | X=1) = 0.05, P(Y=1 | X=1) = 0.95
- Conditional distribution of X given Y:
  - P(X=0 | Y=0) = 0.975, P(X=1 | Y=0) = 0.025
  - P(X=0 | Y=1) = 0.269, P(X=1 | Y=1) = 0.731

We calculated P(X=1 | Y=1) earlier; can you calculate the other conditional probabilities?

- Marginal distributions:
  - P(X=0) = 0.595 + 0.105 = 0.7, P(X=1) = 0.015 + 0.285 = 0.3
  - P(Y=0) = 0.595 + 0.015 = 0.61, P(Y=1) = 0.105 + 0.285 = 0.39

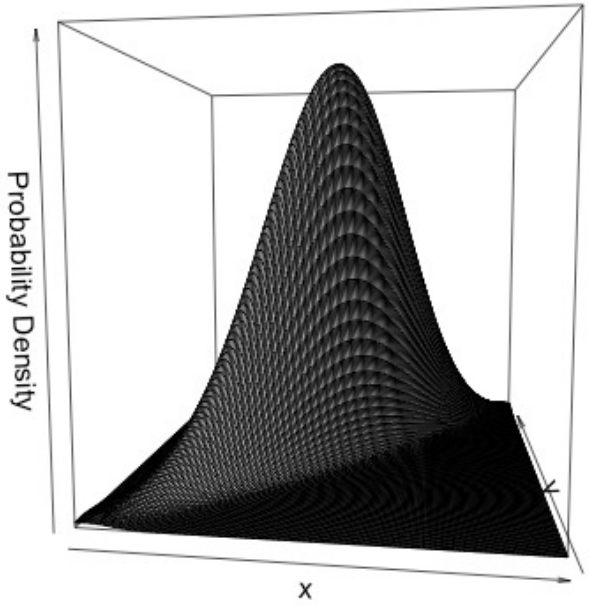
Example Joint Density Functions for 2 Random Variables

Independent Random Variables



$f(x,y)=f(x)f(y)$

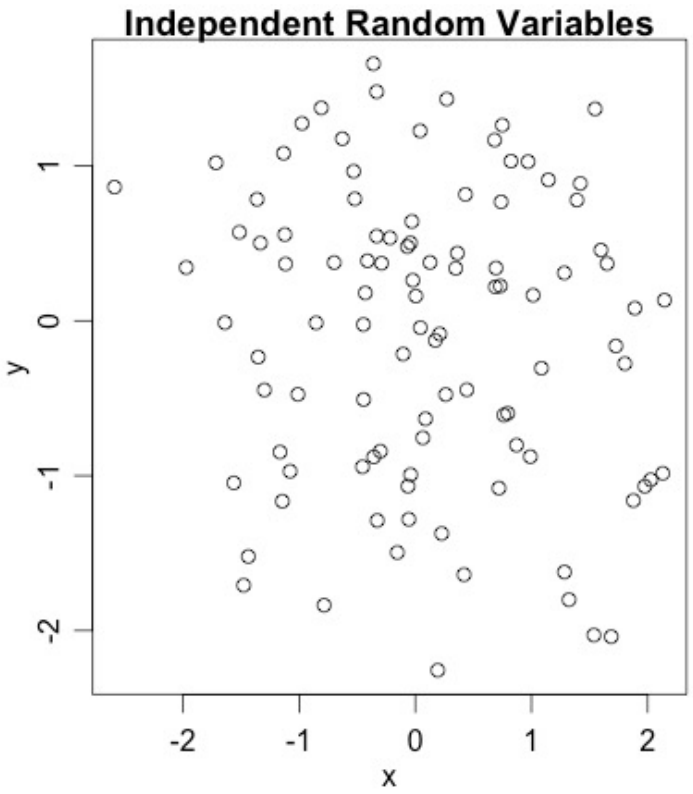
Correlated Random Variables

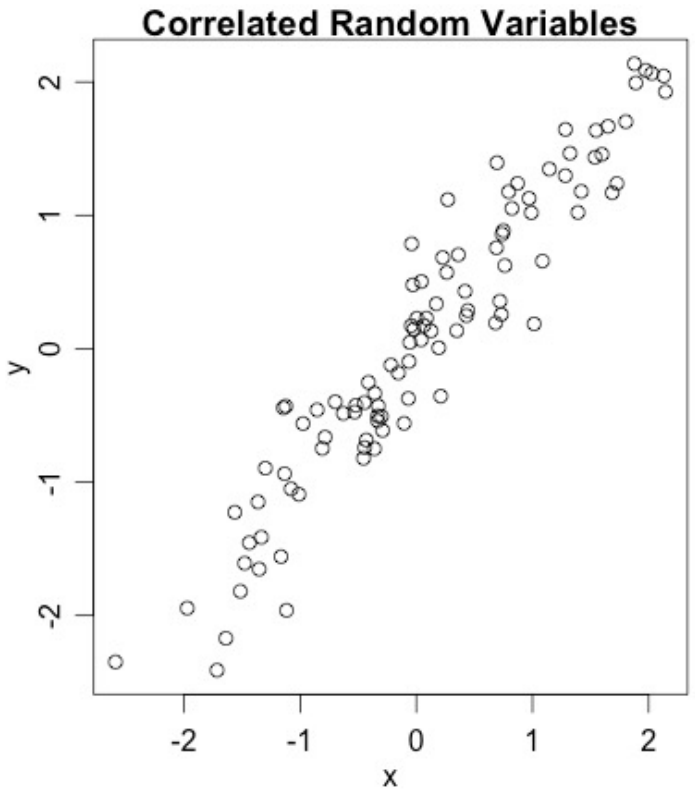


$f(x,y)=f(x)f(y|x)$

R code for generating these plots is provided with this module

Scatterplots: Random Samples from Independent & Correlated Bivariate Distributions





Covariance

- The covariance matrix measures how random quantities vary together
  - Diagonal elements are variances
  - The $(i,j)^{th}$ off-diagonal element is called $Cov(X_i,X_j)$ or covariance of X_i and X_j,

$$\Sigma = Cov(\underline{X}) = \begin{bmatrix} E[(X_1-\mu_1)^2] & \cdots & E[(X_1-\mu_1)(X_n-\mu_n)] \\ \vdots & \ddots & \vdots \\ E[(X_n-\mu_n)(X_1-\mu_1)] & \cdots & E[(X_n-\mu_n)^2] \end{bmatrix} = \begin{bmatrix} Var(X_1) & \cdots & Cov(X_1,X_n) \\ \vdots & \ddots & \vdots \\ Cov(X_1,X_n) & \cdots & Var(X_n) \end{bmatrix}$$

  - If X_i and X_j are independent then $Cov(X_i,X_j) = 0$
    - Covariance matrix is diagonal if random variables are all mutually independent
  - $Cov(X_i,X_j)$ is positive if X_i tends to increase with X_j and negative if X_i tends to decrease with X_j
  - If X_i and X_j have an *exact* [positive / negative] linear relationship then $Cov(X_i,X_j)$ is equal to [plus / minus] the product of the standard deviations of X_i and X_j

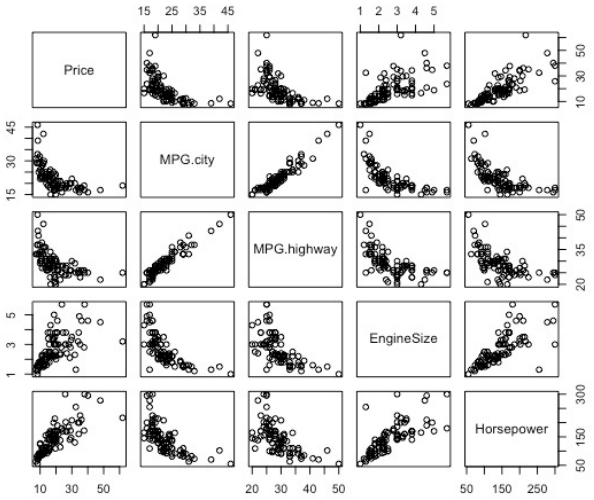
Correlation

- The correlation matrix is defined as: $\text{corr}(\mathbf{X}) = (\text{diag}(\Sigma))^{-\frac{1}{2}} \, \Sigma \, (\text{diag}(\Sigma))^{-\frac{1}{2}}$
  - The diagonal elements are equal to 1
  - The $(i,j)^{th}$ off-diagonal element is the covariance divided by the product of the i^{th} and j^{th} standard deviations

$$Corr(\underline{X}) = \begin{bmatrix} 1 & \cdots & \frac{E[(X_1 - \mu_1)(X_n - \mu_n)]}{\sqrt{E[(X_1 - \mu_1)^2]E[(X_n - \mu_n)^2]}} \\ \vdots & \ddots & \vdots \\ \frac{E[(X_n - \mu_n)(X_1 - \mu_1)]}{\sqrt{E[(X_n - \mu_n)^2]E[(X_1 - \mu_1)^2]}} & \cdots & 1 \end{bmatrix}$$

  - The off-diagonal elements are real numbers between -1 and 1
    - If RVs are independent then correlation is zero
    - The closer the correlation is to 1 or -1, the more nearly linearly related the RVs are

Sample Covariance and Correlation



- Sample covariance matrix
  - Diagonal entry in row i is sample variance of X_i

$$S_i^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ik}-\bar{X}_i)^2$$

  - Off-diagonal element in row i and column j ≠ i is sample covariance of X_i and X_j

$$C_i = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ik}-\bar{X}_i)(X_{jk}-\bar{X}_j)$$

- Sample correlation matrix
  - Diagonal entries are equal to 1
  - Off-diagonal element in row i and column j ≠ i is sample correlation of X_i and X_j

$$R_i = \frac{C_{ij}}{S_i S_j}$$

```
library(MASS)    #Package w many example data sets
data("Cars93")   #Car data set
cor(Cars93[c(5,7,8,12,13)])
```

```
                Price   MPG.city MPG.highway EngineSize Horsepower
Price       1.0000000 -0.5945622  -0.5606804  0.5974254  0.7882176
MPG.city   -0.5945622  1.0000000   0.9439358 -0.7100032 -0.6726362
MPG.highway -0.5606804  0.9439358   1.0000000 -0.6267946 -0.6190437
EngineSize  0.5974254 -0.7100032  -0.6267946  1.0000000  0.7321197
Horsepower  0.7882176 -0.6726362  -0.6190437  0.7321197  1.0000000
```

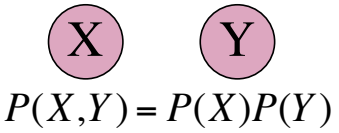
https://www.rdocumentation.org/packages/MASS/versions/7.3-55/topics/Cars93

Independence and Conditional Independence

- $\underline{X}$ and $\underline{Y}$ are *independent* if the conditional distribution of $\underline{X}$ given $\underline{Y}$ does not depend on $\underline{Y}$
  - When $\underline{X}$ and $\underline{Y}$ are independent, the joint density (or mass) function is equal to the product of the marginal density (or mass) functions
  - $f(\underline{x},\underline{y}) = f(\underline{x})f(\underline{y})$
- $\underline{X}$ and $\underline{Y}$ are *conditionally independent* given $\underline{Z}$ if:
  - $f(\underline{x},\underline{y},\underline{z}) = f(\underline{x}|\underline{z})f(\underline{y}|\underline{z})f(\underline{z})$ *General form is* $f(x,y,z) = f(x|y,z)f(y|z)f(z)$
- Conditional independence relationships simplify specificiation of the joint distribution

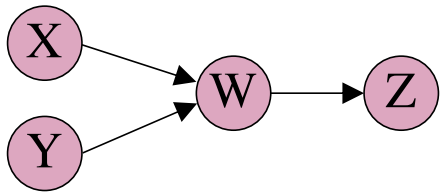
Graphical Representation of Conditional Dependence

- Graphs provide a powerful tool for visualizing and formally modeling dependencies between random variables
  - Random variables (RVs) are represented by nodes
  - Direct dependencies are represented by edges
  - Absence of an edge means no direct dependence



$$P(X,Y) = P(X)P(Y)$$

X and Y are independent



$$P(X,Y,Z,W) = P(X)P(Y)P(W \mid X,Y)P(Z \mid W)$$

X and Y are independent. W depends on X and Y. Z depends on W and is conditionally independent of X and Y given W.

Joint Distribution for Multivariate Random Variable: Summary

- A joint distribution models related random variables
- The marginal distribution models one or more random variable(s) integrating over the other(s)
- The conditional distribution models one or more random variable(s) given the value(s) of the other(s)
- Covariance and correlation matrices measure relationships between random variables (theoretical and sample)
- Scatterplot matrices allow us to visualize relationships in sample data
- Graphical probability models provide a way to visualize conditional dependence relationships

Parametric Families of Distributions

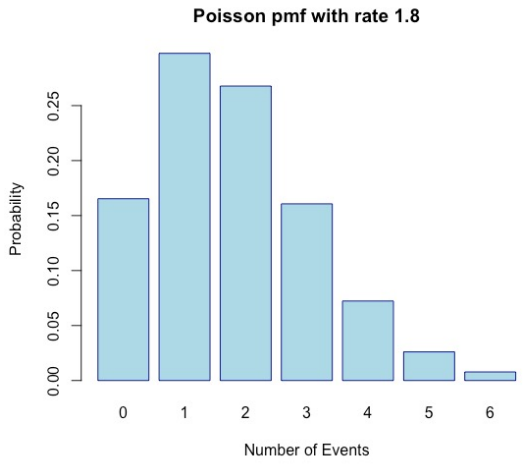
- Statistics makes use of parametric families of distributions
  - Many problems are modeled by assuming observations $X_1, \ldots, X_N$ are independent and identically distributed observations from a distribution with gpdf $f(x \mid \Theta)$
    - Functional form $f(x \mid \Theta)$ is known but value Θ of parameter is unknown
  - Parameter may be a number Θ or a vector $\underline{\Theta} = (\Theta_1, \ldots, \Theta_p)$
- Canonical (standard) problems in statistics:
  - Use sample to construct an estimate (point or interval) of Θ
  - Test a hypothesis about the value of Θ
  - Is the functional form $f(\underline{X} \mid \Theta)$ an adequate model for the data?
  - Predict features (e.g., mean) of a future sample $X_{1+N}, \ldots, X_{M+N}$
- Many distributions have multiple parameterizations in common use and it is important not to confuse different parameterizations

Some Discrete Parametric Families (p. 1 of 2)

- *Binomial distribution* with number of trials *n* and success probability π:
  - Sample space: Non-negative integers (between 0 and *n*)
  - pmf: $f(x \mid n,\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$
  - $E[X|n,\pi] = n\pi;\ Var[X|n,\pi] = n\pi\,(1-\pi)$

Bernoulli distribution is a Binomial distribution with n=1

- *Multinomial distribution* with number of trials *n* and category probability $\underline{\pi}$:
  - Sample space: Vectors of non-negative integers that sum to *n*
  - pmf: $f(x_1 \ldots, x_p \mid n, \pi_1, \ldots, \pi_p) = \left(\frac{n!}{x_1! \cdots x_p!}\right) \prod_{i=1}^{p} \pi_i^{x_i}$
  - $E[\underline{X}|n,\underline{\pi}] = n\underline{\pi};\ Var[X_i|n,\pi_i] = n\pi_i(1-\pi_i)$
- *Poisson distribution* with rate λ:
  - Sample space: Non-negative integers
  - pmf: $f(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$
  - $E[X \mid \lambda] = \lambda\,;\ Var[X \mid \lambda] = \lambda$



Some Discrete Parametric Families (p. 2 of 2)

- *Negative binomial distribution* with size α and probability π:
  - Sample space: Non-negative integers
  - pmf: $f(x \mid \alpha, \pi) = \binom{\alpha + x - 1}{\alpha} \pi^{\alpha} (1-\pi)^{x}$
  - $E[X|\alpha,\pi] = \frac{\alpha(1-\pi)}{\pi}$; $V[X|\alpha,\pi] = \frac{\alpha(1-\pi)}{\pi^2}$

Distribution for number of failures in a sequence of independent and identically distributed trials until α successes occur

- *Beta-binomial distribution* with shape parameters α, β and number of trials *n* (or probability parameter $p = \frac{\alpha}{\alpha+\beta}$, overdispersion parameter $m = \alpha + \beta$ and number of trials n)
  - Sample space: Non-negative integers (between 0 and *n*)
  - pmf: $f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{n}{x} \frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}$
  - $E[X|\alpha,\beta,n] = \frac{n\alpha}{\alpha+\beta} = np$; $Var[X|\alpha,\beta,n] = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)} = np(1-p)\frac{(m+n)}{m+1}$

Gamma function
$\Gamma(y) \equiv \int_0^{\infty} u^{y-1} e^{-u} du$

Some Continuous Parametric Families (p. 1)

- *Gamma distribution* with shape α and scale β:
  - Sample space: Nonnegative real numbers
  - pdf: $f(x \mid \alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$
  - $E[X \mid \alpha,\beta] = \alpha\beta$; $Var[X \mid \alpha,\beta] = \alpha\beta^2$

Gamma distribution is also parameterized with shape and rate r = 1/β

Exponential distribution is a Gamma distribution with α=1

Chi-square distribution with δ degrees of freedom is a Gamma distribution with α=δ/2 and β=2

- *Beta distribution* with shape parameters α and β:
  - Sample space: Real numbers between 0 and 1
  - pdf: $f(x \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
  - $E[X \mid \alpha,\beta] = \alpha/(\alpha+\beta)$; $Var[X \mid \alpha,\beta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Uniform distribution is a Beta distribution with α=β=1

Gamma function

$$\Gamma(y) \equiv \int_0^{\infty} u^{y-1} e^{-u} du$$

- Univariate *normal distribution* with mean μ and standard deviation σ:
  - Sample space: Real numbers
  - pdf: $f(x \mid \mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\}$
  - $E[X \mid \mu,\sigma] = \mu$; $Var[X \mid \mu,\sigma] = \sigma^2$

Some Continuous Parametric Families (p. 2)

- *Multivariate normal distribution* with mean $\underline{\mu}$ covariance matrix Σ:
  - Sample space: Vectors of real numbers
  - pdf: $f(\underline{x} \mid \underline{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi \mid \Sigma \mid}} \exp\left\{-\frac{1}{2}\left(\underline{x} - \underline{\mu}\right)' \Sigma^{-1} \left(\underline{x} - \underline{\mu}\right)\right\}$
  - $E\left[\underline{X} \middle| \underline{\mu}, \underline{\Sigma}\right] = \underline{\mu};\ Var\left[\underline{X} \middle| \underline{\mu}, \underline{\Sigma}\right] = \underline{\Sigma}$
- *Dirichlet distribution* with shape parameters $\alpha_1, \alpha_2, \ldots, \alpha_p$:
  - Sample space: Real non-negative vectors summing to 1
  - pdf: $f(x_1, \ldots, x_p \mid \alpha_1, \ldots, \alpha_p) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_p)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_p)} x_1^{\alpha_1 - 1} \cdots x_p^{\alpha_p - 1}$
  - $E[X_i \mid \alpha_1, \ldots, \alpha_p] = \frac{\alpha_i}{\sum_j \alpha_j} \quad Var[X_i \mid \alpha_1, \ldots, \alpha_p] = \frac{\alpha_i (1 - \alpha_i)}{\left(\sum_j \alpha_j\right)^2 \left(\sum_j \alpha_j + 1\right)}$
  - Dirichlet distribution is a multivariate generalization of the Beta distribution

Sufficient Statistic

- A sufficient statistic is a data summary (function of sample of observations) such that the observations are independent of the parameter given the sufficient statistic
  - Example: For a sample of n iid observations from a Poisson distribution, the sum of the observations is a sufficient statistic for the rate parameter
  - Example: For a sample of n observations from a Bernoulli distribution, the total number of successes is a sufficient statistic for the success probability
- If observations $X_1, \ldots, X_n$ are sampled randomly from a distribution with gpdf $f(X \mid \theta)$ and Y is sufficient for the parameter θ, then the posterior distribution $f(\theta|X)$ depends on the observations only through the sufficient statistic
  - A sufficient statistic contains all information needed to calculate the posterior distribution for θ
- Fisher's factorization theorem: $T(X)$ is sufficient for θ if and only if the conditional probability distribution $f(X \mid \theta)$ can be factored as:

$$f(X \mid \theta) = h(x)g_\theta(T(x))$$

Distributions in R

Distribution	Base name	Parameters
beta	beta	shape1, shape2
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-squared	chisq	df
exponential	exp	rate
F	f	df1, df2
gamma	gamma	shape, rate
geometric	geom	p
hypergeometric	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
logistic	logis	location, scale
negative binomial	nbinom	size, prob
normal	norm	mean, sd
Poisson	pois	lambda
Student t	t	df
uniform	unif	min, max
Weibull	weibull	shape, scale

In R, there are four functions for each distribution, invoked by adding a prefix to the distribution's base name:

- p for "probability" – the cumulative distribution function (cdf)
- q for "quantile" – the inverse cdf
- d for "density" – the density or mass function
- r for "random" – generates random numbers from the distribution

Examples: rpois, dgamma

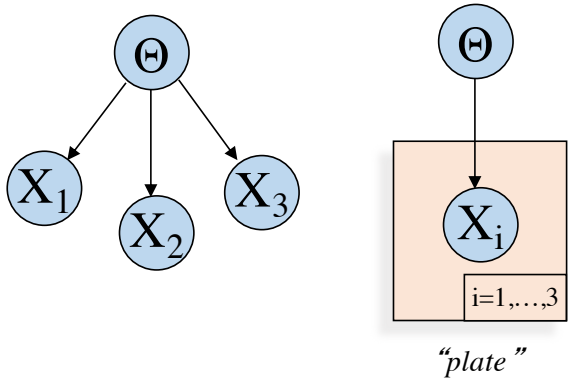
Source:
http://www.johndcook.com/distributions_R_SPLUS.html

- *Information on distributions in R:*
  http://www.stat.umn.edu/geyer/old/5101/rlook.html
- *Some distributions we will use are not provided in base R but are available in R packages*

Independent and Identically Distributed Observations: Plate Representation

- Statistical models often assume the observations are a random sample from a parameterized distribution
- Mathematically, this is represented as independent and identically distributed (iid) conditional on the parameter Θ
- The gpdf for an iid sample $X_1, \ldots, X_n$ conditional on Θ is written as a product of factors:

$$f(x_1, x_2, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) f(x_2 \mid \theta) \cdots f(x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$



"plate"

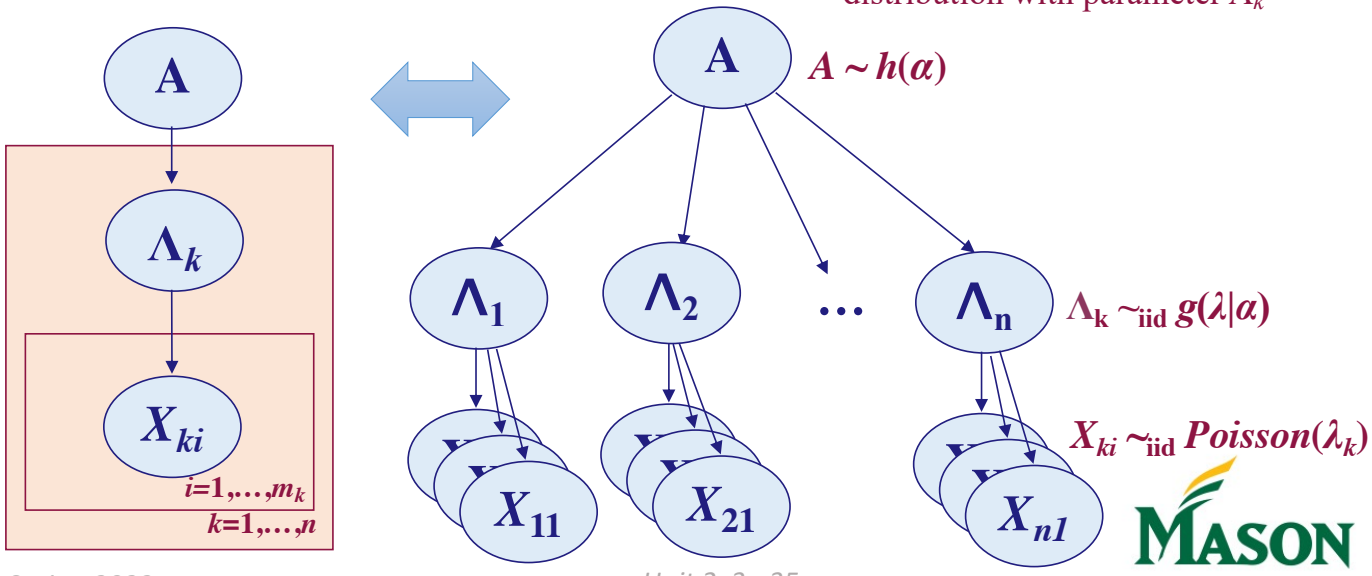
- *A "plate" represents repeated structure*
- *The RVs inside the plate are iid conditional on their parents*
- *In this example, the X_i are iid given Θ*
- ***Joint** gpdf for ($\underline{X}$,Θ):*

  $g(\theta)\prod_i f(x_i \mid \theta)$
- ***Conditional** gpdf for $\underline{X}$ given Θ:*

  $\prod_i f(x_i \mid \theta)$

Multi-Level Models: Graphical Representation

- We can use graphical probability models to:
  - Specify complex multivariate distributions compactly
  - Visualize dependence relationships, and
  - Perform inference efficiently
- Parametric distributions are often used as building blocks to construct complex graphical probability models



- Birth rates Λ_k for hospitals k=1, …, n are sampled randomly from a distribution with parameter A
- Hourly births X_{ki} at hospital k are sampled from a Poisson distribution with parameter Λ_k

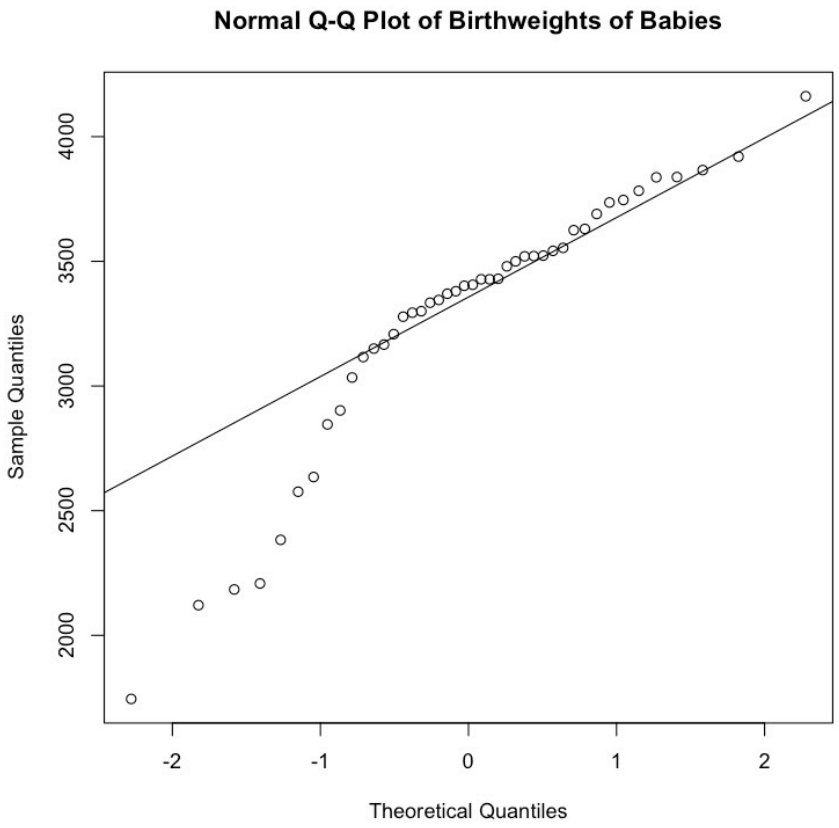
Is a Parametric Distribution a Good Model?

- Before applying a parametric model, we should assess its adequacy
  - Theoretical assumptions underlying the distribution
  - Exploratory data analysis
  - Formal goodness-of-fit tests
- In your homework I expect you to assess whether the parametric model I give you is a good one (sometimes it won't be!)

Some Tools for Exploratory Data Analysis

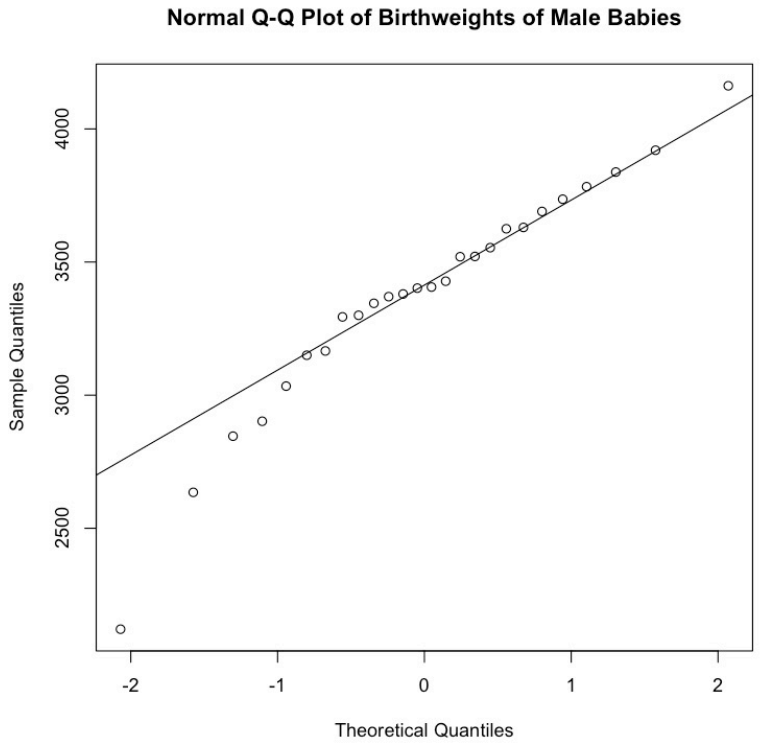
- A q-q plot is a commonly used diagnostic tool
  - Plot quantiles of data distribution against quantiles of theoretical distribution
  - If theoretical distribution is correct, the plot should look approximately like the plot of $y=x$
  - Problem: what about unknown parameters?
  - Solution: We can estimate parameters from data
  - Solution: in the case of a location-scale family, we can plot data quantiles against standard distribution (location parameter = 0 and scale parameter = 1)
    - If theoretical distribution is correct the plot should look approximately like a straight line
    - Slope and intercept of line are determined by scale and location parameters
- Another diagnostic tool is to compare empirical and theoretical counts for discrete RV (or discretized continuous RV)
- What other checks for model fit have you encountered?

Are Birth Weights Normally Distributed?



Data on birth weights of babies born in a Brisbane hospital on December 18, 1997
https://rdrr.io/cran/UsingR/man/babyboom.html

Birth Weights of Boys and Girls

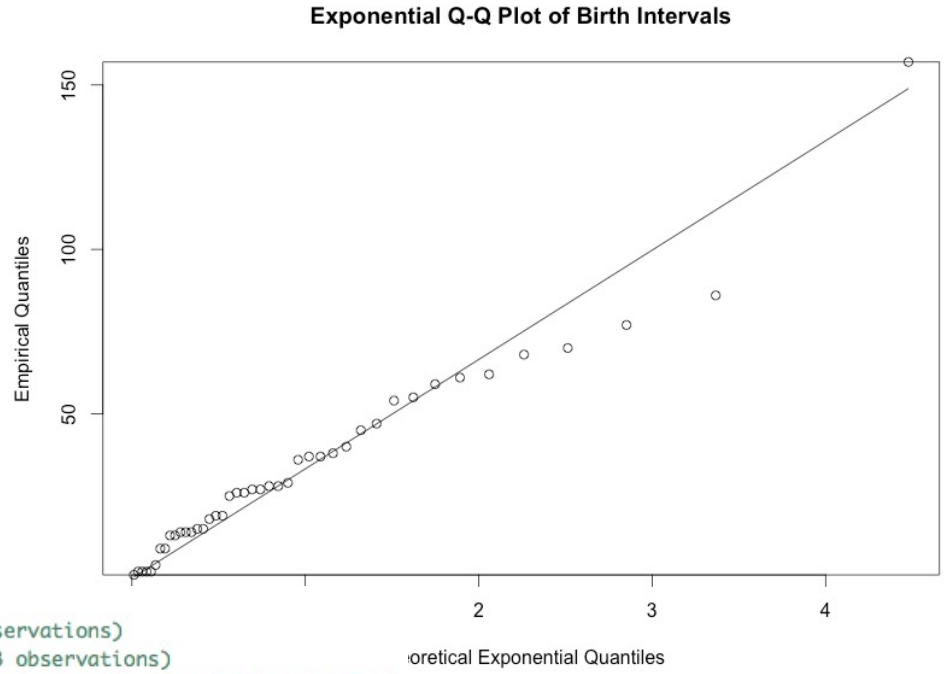


Are Times Between Births Exponentially Distributed?

The data: times between births of babies born in a Brisbane hospital on December 18, 1997

R notes:

- `ppoints` function computes probabilities for evaluating quantiles
- `diff` function computes lagged differences
- `lines` function adds lines to a plot
- `qexp` function computes exponential quantiles

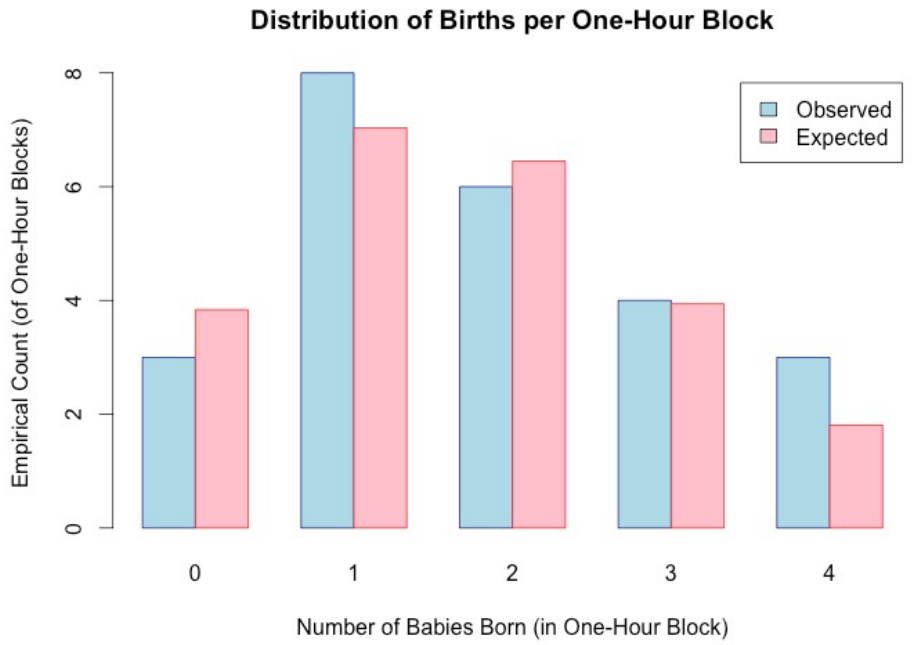



```
# Exponential q-q plot of birth intervals
birthtime=babyboom[,4]          # Time of birth in minutes after midnight (44 observations)
birthinterval=diff(birthtime)   # compute birth intervals using diff function (43 observations)
exponential.quantiles = qexp(ppoints(length(birthtime)))  # quantiles of standard exponential distribution (rate=1)
qqplot(exponential.quantiles, birthinterval,main="Exponential Q-Q Plot of Birth Intervals",
       xlab = "Theoretical Exponential Quantiles", ylab = "Empirical Quantiles")
lines(exponential.quantiles,exponential.quantiles*mean(birthinterval)) # Overlay a line
```

Do Births Per Hour Follow a Poisson Distribution?

Comparing empirical to theoretical counts is a useful tool for evaluating fit (if the expected counts are not too small)

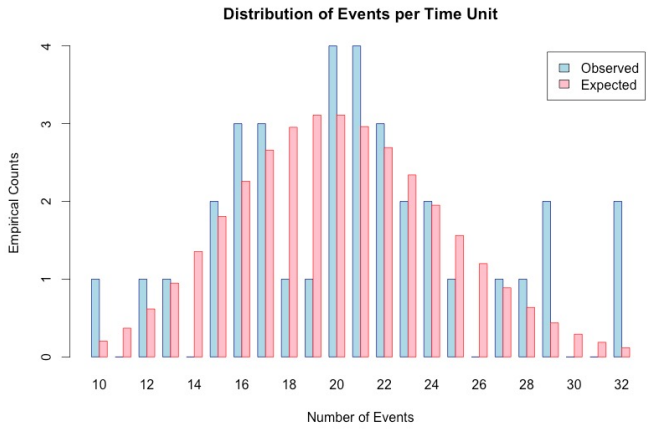
Births	Empirical Count	Expected Count
0	3	3.84
1	8	7.03
2	6	6.45
3	4	3.94
4	3	1.81
5+	0	0.93
TOTAL	24	24.00

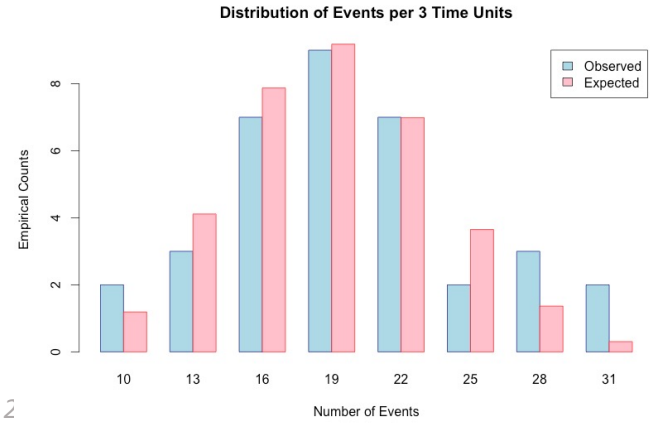


If times (in hours) between births are a random sample from an exponential distribution, then counts of births per hour are a random sample from a Poisson distribution

Frequency Plot is Misleading when Expected Counts are Too Small

- Plots show empirical (blue) and expected (pink) counts from sample of 35 observations from Poisson distribution with mean 20
  - Top plot – bin size is 1 unit
  - Bottom plot – bin size is 3 units
- When expected counts are small, plots of observed and expected counts will not look similar
- Common rule of thumb: choose bin size so most bins have expected count of at least 5



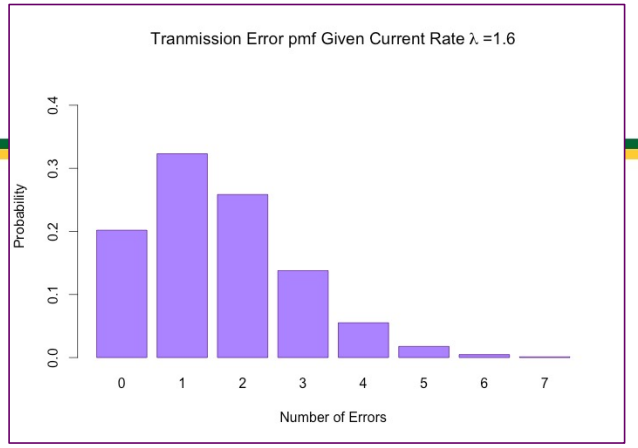


Example: Modeling Transmission Errors

- Number of transmission errors per hour is distributed as Poisson with parameter λ

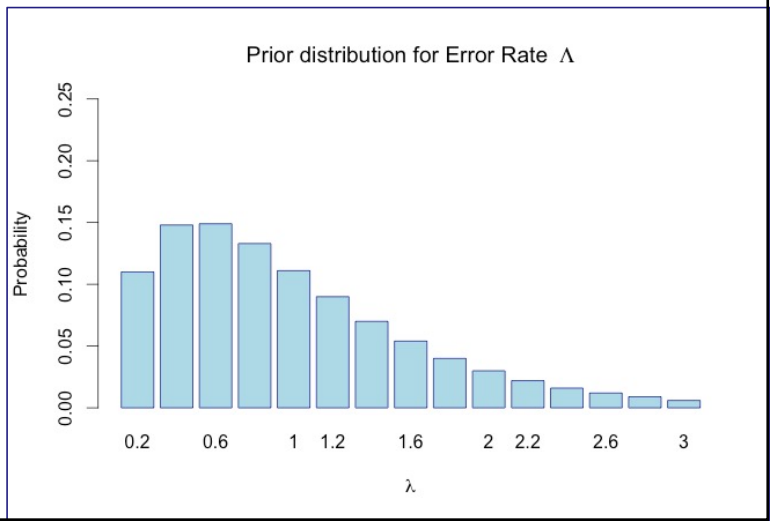
$$f(x \mid \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

- Data on previous system established error rate as 1.6 errors per hour
- New system has design goal of cutting this rate in half to 0.8 errors per hour
- Observe new system for 6 one-hour periods:
  - Data: 1, 0, 1, 2, 1, 0
- Questions:
  - Have we met the design goal?
  - Does the new system improve the error rate?



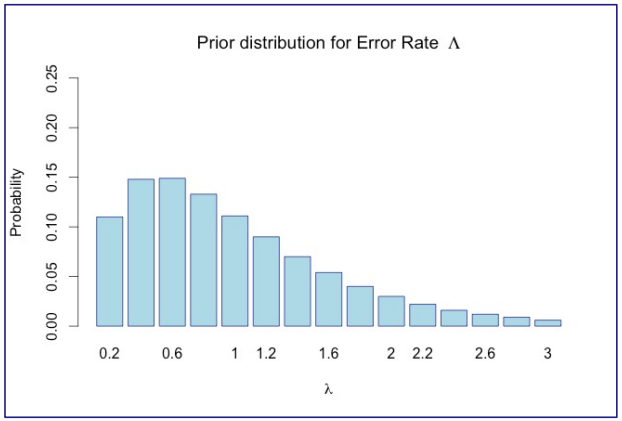
The Prior Distribution (discretized)

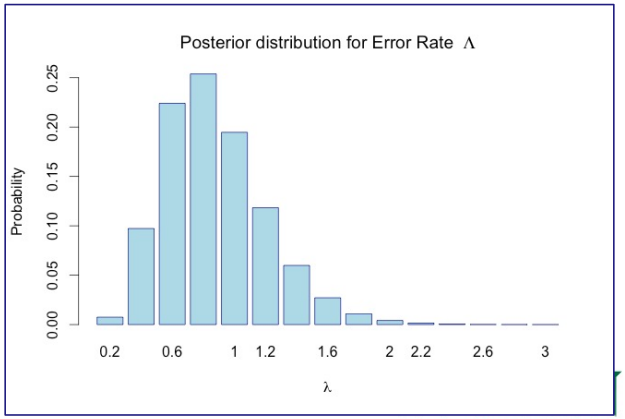
- We use expert judgment to define prior distribution on a discrete set of values
  - Error rate can be any positive real number
  - Later we will revisit this problem with a continuous prior distribution)
- Experts familiar with the new system design said:
  - *"Meeting the design goal of 0.8 errors per hour is about a 50-50 proposition."*
  - *"The chance of making things worse than current rate of 1.6 errors per hour is small but not negligible"*
- Expert agrees that the discretized distribution shown here is a good reflection of his prior knowledge
  - Expected value is about 1.0
  - Distribution is heavy tailed on the right
  - $P(\Lambda \leq 0.8) = 0.54$
  - $P(\Lambda \leq 1.6) = 0.87$
  - Values of Λ greater than 3 are unlikely enough to ignore



Bayesian Inference for Error Rate

```r
# Discretized parameter values (15 equally spaced values)
lambda <- seq(length=15, from=0.2, to=3)  # Grid of values for parameter lambda

# Prior distribution on transmission error rate
priorDist <- c(0.110, 0.148, 0.149, 0.133, 0.111, 0.090, 0.070,
               0.054, 0.040, 0.030, 0.022, 0.016, 0.012, 0.009, 0.006)

# The observations
errors <- c(1,0,1,2,1,0)

# The likelihood function is a product of Poisson pmfs
lik <- array(1,length(lambda))          # Initialize likelihood as a constant
for (i in 1:6) {
  lik <- lik*dpois(errors[i],lambda)    # Multiply by Likelihood
}

# The posterior distribution
postDist <- priorDist*lik/sum(priorDist*lik)
```



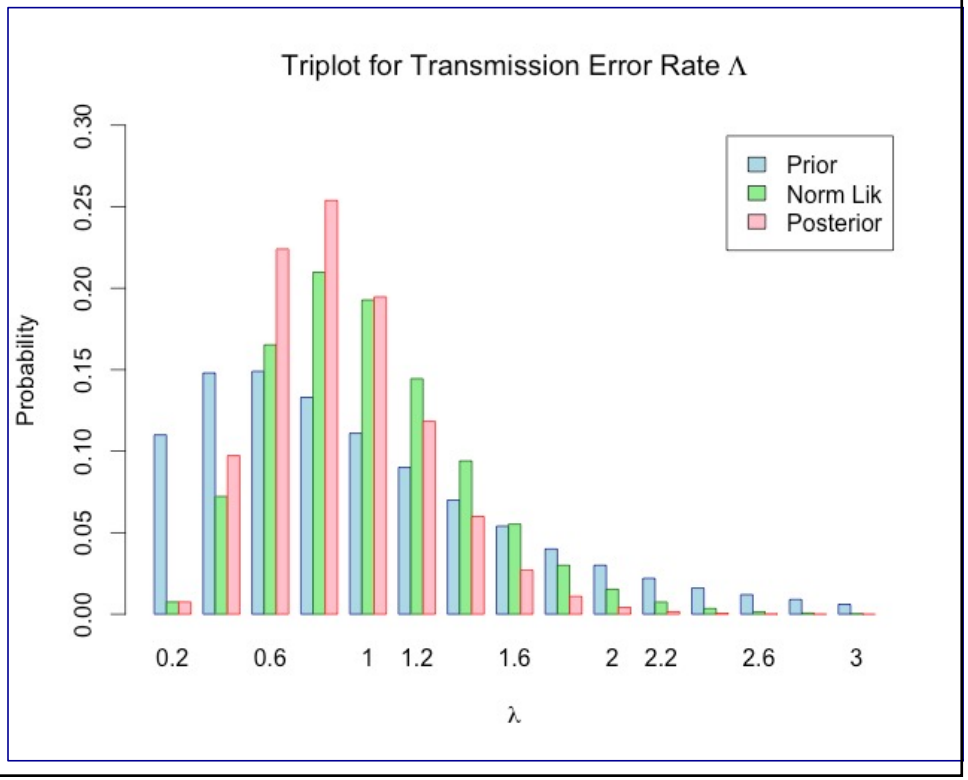


Features of the Posterior Distribution

- Central tendency
  - Posterior mean of Λ is 0.87
  - Prior mean of Λ is 0.97; data mean is .83
  - Typically posterior central tendency is a compromise between the prior distribution and the center of the data
- Variation
  - Posterior standard deviation of Λ is about .33
  - Prior standard deviation of Λ is about .62
  - Typically variation in the posterior is less than variation in the prior (we have more information)
- Meeting the threshold
  - Posterior probability of meeting or doing better than design goal is about .58
  - Posterior probability that new system is better than old system is about .96
  - Posterior probability that new system is worse than old system is less than .02

Triplot

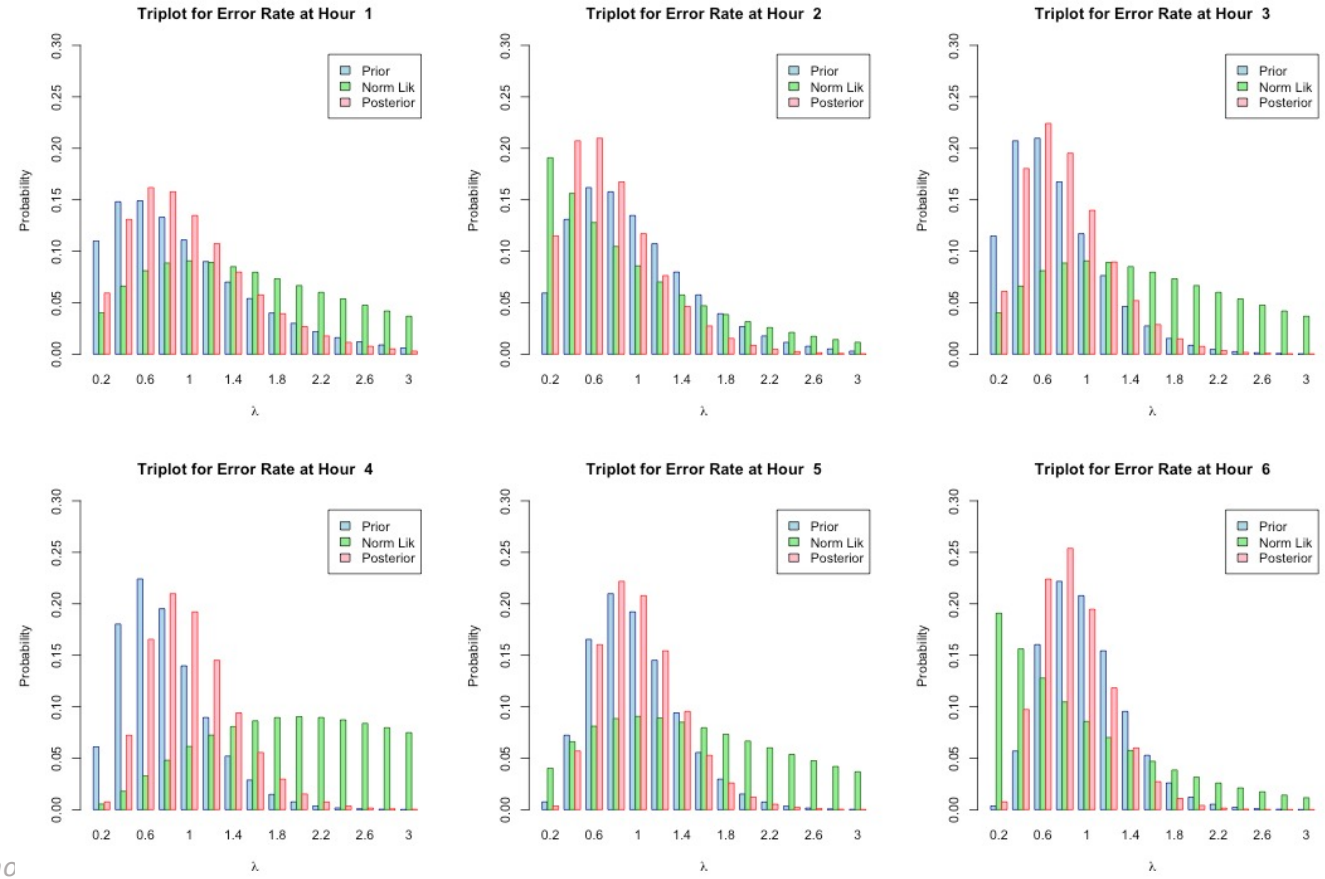
- Visual tool for examining Bayesian belief dynamics
- Plot prior distribution, normalized likelihood, and posterior distribution
- Normalized likelihood:
  - Posterior distribution we would obtain if all values were assigned equal prior probability
  - To calculate, divide likelihood by sum or integral over λ



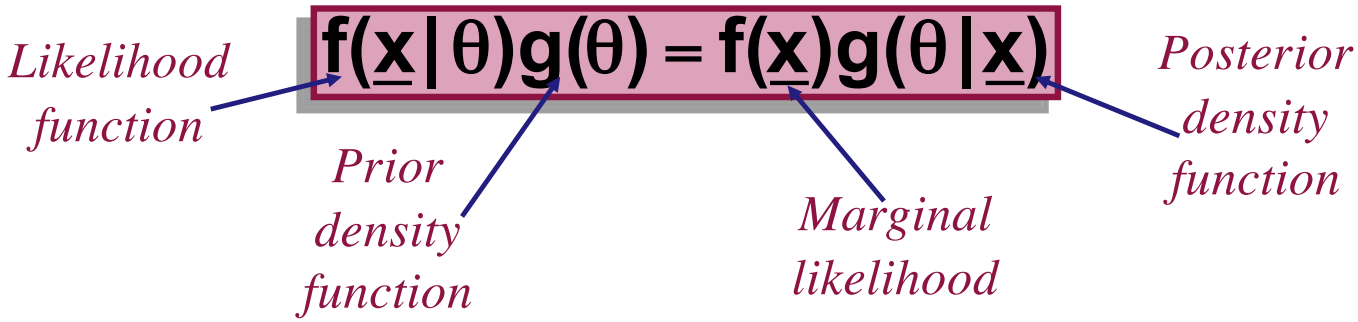
Bayesian Belief Dynamics: Sequential and Batch Processing of Observations

- Batch processing:
  - Use Bayes rule with prior $g(\theta)$ and combined likelihood $f(X_1, \ldots, X_n|\theta)$ to find posterior $g(\theta|X_1, \ldots, X_n)$
- Sequential processing:
  - Use Bayes rule with prior $g(\theta)$ and likelihood $f(X_1|\theta)$ to find posterior $g(\theta|X_1)$
  - Use Bayes rule with prior $g(\theta|X_1)$ and likelihood $f(X_2|\theta)$ to find posterior $g(\theta|X_1, X_2)$
  - …
  - Use Bayes rule with prior $g(\theta|X_1, \ldots, X_{n-1})$ and likelihood $f(X_n|\theta)$ to find posterior $g(\theta|X_1, \ldots, X_n)$
- The posterior distribution after n observations is the same with both methods

Visualizing Bayesian Belief Dynamics: Hour-by-Hour Triplots for Transmission Error Data



Fundamental Identity of Bayesian Inference



- The joint gpdf of parameter and data can be expressed in two ways:
  - Prior density for parameter times likelihood function
  - Marginal likelihood times posterior density for parameter
    - Marginal likelihood is (conditional) likelihood integrated over parameter

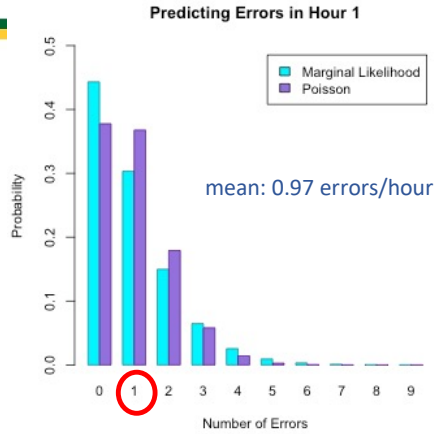
$$f(\underline{x}) = \int_\theta f(\underline{x}|\theta)g(\theta)d\mu(\theta) = \begin{cases} \int_\theta f(\underline{x}|\theta)g(\theta)d\theta & \text{continuous parameter} \\ \sum_\theta f(\underline{x}|\theta)g(\theta) & \text{discrete parameter} \end{cases}$$

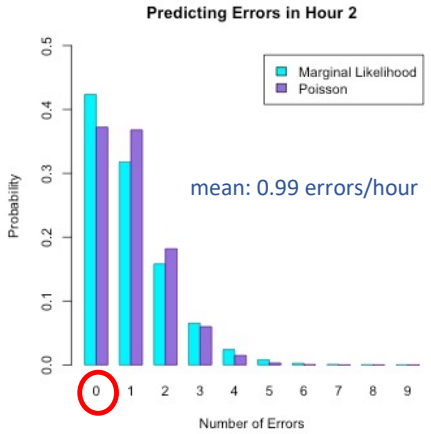
Marginal Likelihood

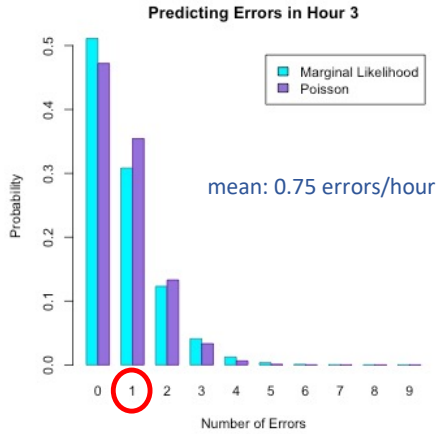
- ***Before we have seen X***, we use the marginal likelihood to predict the value of X
  - When used for predicting X, the marginal likelihood is called the *predictive distribution* for X
- ***After we see X=x***, we divide the joint probability $f(x|\theta)g(\theta)$ by the marginal likelihood $f(x)$ to obtain the posterior probability mass function of θ:
  - $g(\theta \mid x) = \dfrac{f(x \mid \theta)g(\theta)}{f(x)}$
  - The marginal likelihood $f(x)$ is the normalizing constant in Bayes Rule – we divide by $f(x)$ to ensure that the posterior probabilities sum to 1
- The marginal likelihood $f(x) = \sum_\theta f(x|\theta)g(\theta)$ includes uncertainty about both θ and x given θ
- Non-Bayesians sometimes predict future observations using a point estimate of θ
- Predictions using the marginal likelihood are more spread out (include more uncertainty) than predictions using a point estimate of θ

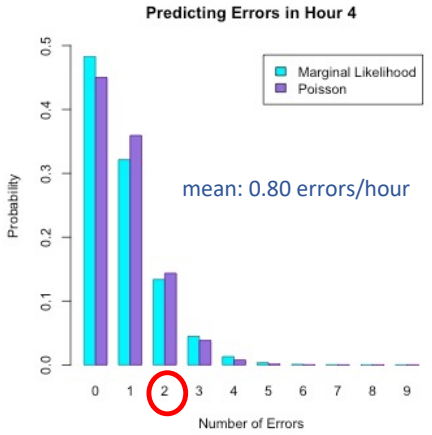
Predicting Future Observations:

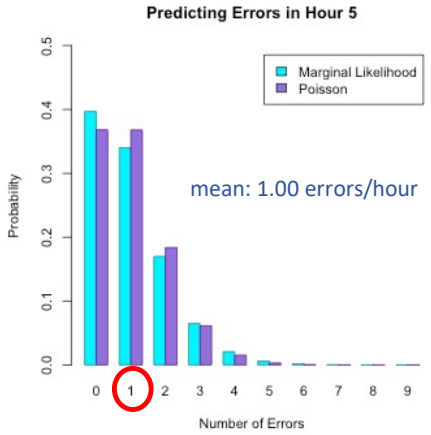
Compare Marginal Likelihood with Poisson(E[Λ | observations so far])

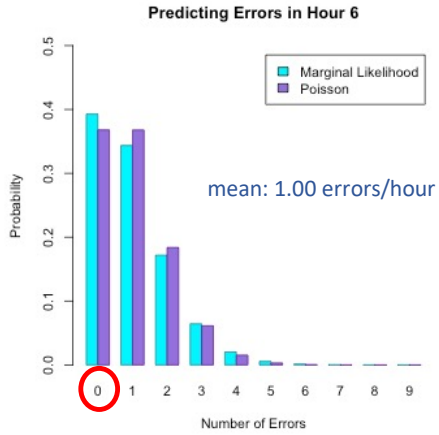












Parameters and Conditioning: Frequentists and Subjectivists

- Frequentist view of parameters
  - Parameter θ represents a true but unknown feature of a random data generating process
  - It is not appropriate to put a probability distribution on θ because the value of θ is not random
- Subjectivist view of parameters
  - A subjectivist puts a probability distribution on θ as well as X_i given θ
  - Parametric distributions are a convenient means to specify probability distributions that represent our beliefs about as yet unobserved data
  - Many subjectivists don't believe "true parameters" exist
- Frequentists and subjectivists on conditioning
  - Frequentists condition on parameter and base inferences on data distribution $f(x_1|\theta) \cdots f(x_n|\theta)$
    - Even after X has been observed it is treated as random
  - Subjectivists condition on knowns and treat unknowns probabilistically
    - Before observing data $X = x$, the joint distribution of parameters and data is $f(x_1|\theta) \cdots f(x_n|\theta)g(\theta)$
    - After observing data, $X = x$ is known and random variable Θ has distribution $g(\theta|x_1, \ldots, x_n)$

Connecting Subjective and Frequency Probability: de Finetti's Representation Theorem

IF

- Your beliefs are represented by a probability distribution P over a sequence of events $X_1, X_2, \ldots$
- You believe the sequence is *infinitely exchangeable* (your probability for any sequence of successes and failures does not depend on the order of successes and failures)

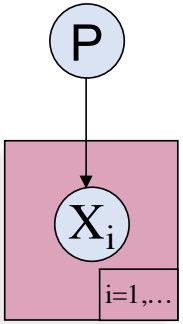
THEN

- You believe with probability 1 that the proportion of successes will tend to a definite limit as the number of trials goes to infinity

$$\frac{S_n}{n} \rightarrow p \text{ as } n \rightarrow \infty$$ where S_n = # successes in n trials

- Your probability distribution for S_n is the same as the distribution you would assess if you believed the observations were random draws from some unknown "true" value of p with probability density function $f(p)$

$$P(S_n = k) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} g(p) dp$$



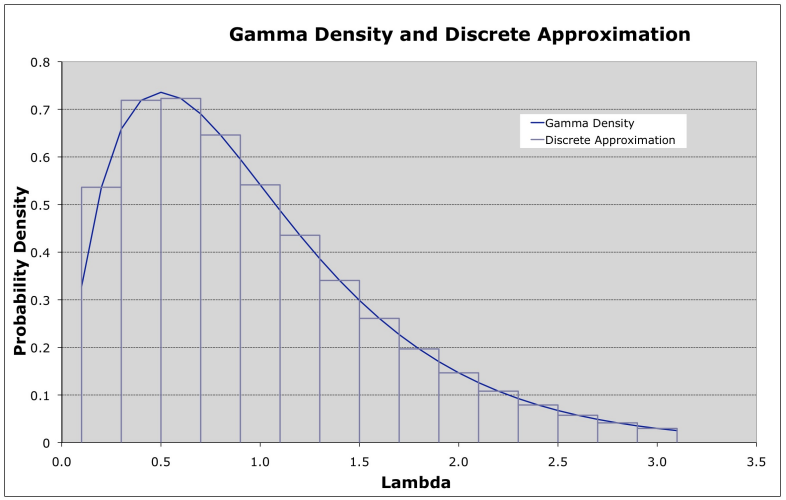
A sequence a frequentist would call random draws from a "true" distribution is one a Bayesian would call exchangeable

Bayesian Inference for Continuous Random Variables

Inference for continuous random variable is limiting case of inference with discretized random variable as number of bins and width of bin goes to zero

- Accuracy tends to increase with more bins
- Accuracy tends to increase with smaller area per bin
- Be careful with tail area of unbounded random variables
- Closed form solution for continuous problem exists in special cases (see Unit 3)

- *The prior distribution we used for the transmission error example was obtained by discretizing a Gamma distribution*
- *In Unit 3, we will compare with the exact result using the Gamma prior distribution*



Summary and Synthesis

- A random variable represents an uncertain hypothesis
  - Categorical, ordinal, discrete numerical, continuous numerical
  - Function from sample space to outcomes (usually real numbers)
  - Used to define events
- Probability mass functions, density functions, and cumulative distribution functions are tools for defining probability distributions
  - We examined measures of central tendency and spread
- Parametric families of distributions are convenient and practical "off the shelf" models for common types of uncertain processes
  - We listed several commonly used parametric families
  - We noted that observations are often modeled as randomly sampled from a parametric family with unknown parameter
  - We applied Bayes rule to infer a posterior distribution for a parameter (using a discretized prior distribution)
  - We contrasted the subjectivist and frequentist approaches to parameter inference
  - We saw that de Finetti's exchangeability theorem provides a connection between the frequentist and subjectivist views of parameters