A Brief Introduction to Probability and Statistics

Yu Wang

Jan 13, 2019

UNIVERSITY OF CALGARY

OUTLINE

01
Why Statistics?

02
Introduction to Probability

03
Introduction to Statistics

04
An Example

Why Statistics?

People who play Statistics

People who played by Statistics

UNIVERSITY OF CALGARY

Why Statistics?

❑Predict The Stock Market

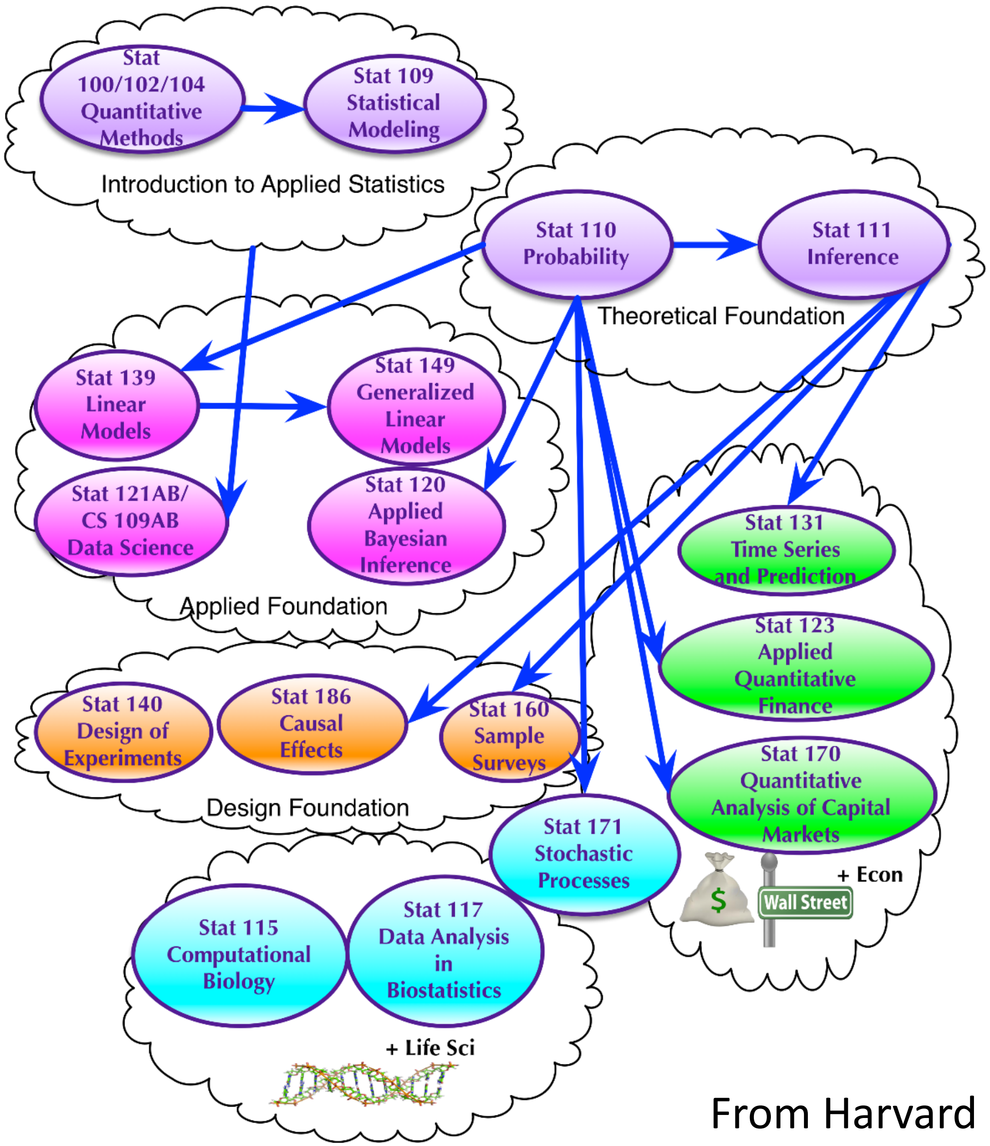
Suppose that one Monday morning you receive in the mail a letter from a firm with which you are not familiar, stating that the firm sells forecasts about the stock market for very high fees. To indicate the firm's ability in forecasting, it predicts that a particular stock, or a particular portfolio of stocks, will rise in value during the coming week. If they successfully predict 7 consecutive weeks, Are you going to pay them to buy another predicts?

The probability of Successfully predict for 7 weeks $(1/2)^7 = 0.008$

Where do we use statistics?

- Scientific Research
- Weather Forecasting
- Insurance
- Quantitative Trading
- Medicine
- Disease prediction
- Quality test
- Election
- …

What included in Statistics



From Harvard University

UNIVERSITY OF CALGARY

PART1 Introduction to Probability

Probability

❑ Contents of Probability

- Combinatorial Analysis
- Axioms of Probability
- Conditional Probability
- Random Variables
- Expectation and Variance
- Special Distributions
- The Law of Large Numbers and Central Limit Theorem
- Generating Functions
- Markov Chains
- Random Walks

Probability

- Sample Space
  - The set of all possible outcomes of an experiment
- Events
  - Any subset E of the sample space

[EXAMPLE]

If the experiment consists of flipping two coins, then the sample space consists of the following four points:

$$S = \{(H,H), (H, T), (T,H), (T, T)\}$$

if $E = \{(H,H), (H, T)\}$, then E is the event that a head appears on the first coin

Probability

Definition of Probability

- We suppose that an experiment, whose sample space is *S*, is repeatedly performed under exactly the same conditions. For each event *E* of the sample space *S*, we define *n(E*) to be the number of times in the first n repetitions of the experiment that the event *E* occurs. Then *P(E)*, the probability of the event *E*, is defined as

$$P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$$

P(E) is defined as the (limiting) proportion of time that *E* occurs

UNIVERSITY OF CALGARY

Probability

❑ Random Variables

- Let S be the sample space for an experiment. A real-valued function that is defined on S is called a *random variable*
- Including Discrete R.V. and Continuous R.V.
- we may assign probabilities to the possible values of the random variable

[EXAMPLE]

Suppose that our experiment consists of tossing 3 fair coins. If we let Y denote the number of heads that appear, then Y is a random variable taking on one of the values 0, 1, 2, and 3 with respective probabilities

$$P\{Y=0\} = P\{(T,T,T)\} = \frac{1}{8}$$
$$P\{Y=1\} = P\{(T,T,H),(T,H,T),(H,T,T)\} = \frac{3}{8}$$
$$P\{Y=2\} = P\{(T,H,H),(H,T,H),(H,H,T)\} = \frac{3}{8}$$
$$P\{Y=3\} = P\{(T,T,T)\} = \frac{1}{8}$$

UNIVERSITY OF CALGARY

Probability

❑ Distribution

- Let X be a random variable. The distribution of X is the collection of all probabilities of the form Pr(X ∈ C) for all sets C of real numbers such that {X ∈ C} is an event.
- Probability mass function(P.M.F) and Probability density function(P.D.F)

[EXAMPLE]

$p(x) = c\lambda^x/x!,\ x = 0, 1, 2, \ldots,$ Poisson distribution

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normal distribution

Probability

Expectation

- The expectation of X is a weighted average of the possible values that X can take on, each value being weighted by the probability that X assumes it

Discrete R.V.

$$E[X] = \sum_{x:p(x)>0} xp(x)$$

Continuous R.V.

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx$$

[EXAMPLE]

Bernoulli distribution E(X)=p; Normal distribution E[X]=μ

Variance

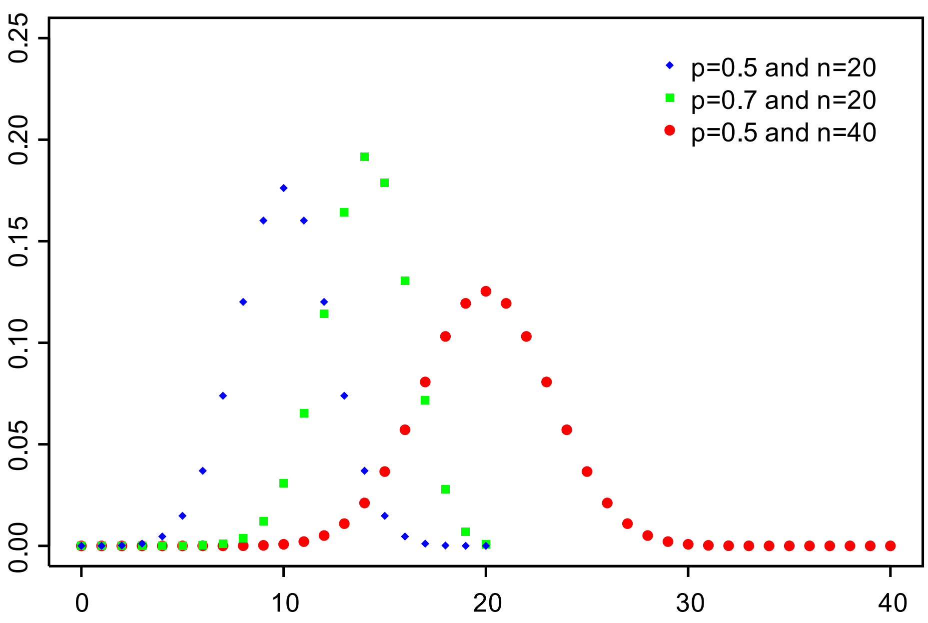
- Let X be a random variable with finite mean $\mu = E(X)$.The *variance of X*, denoted by Var(X), is defined as follows:

$$\mathrm{Var}(X) = E[(X - \mu)^2].$$

$$Var(X) = E(X^2) - [E(X)]^2.$$

Binomial distribution

- Suppose now that n independent trials, each of which results in a success with probability p and in a failure with probability 1 − p, are to be performed. If X represents the number of successes that occur in the n trials, then X is said to be a binomial random variable with parameters (n, p). **X~B(*n, p*)**



PMF

$$\binom{n}{k} p^k (1-p)^{n-k}$$

Expectation

$$E(x)=np$$

Variance

$$Var(x)=np(1-p)$$

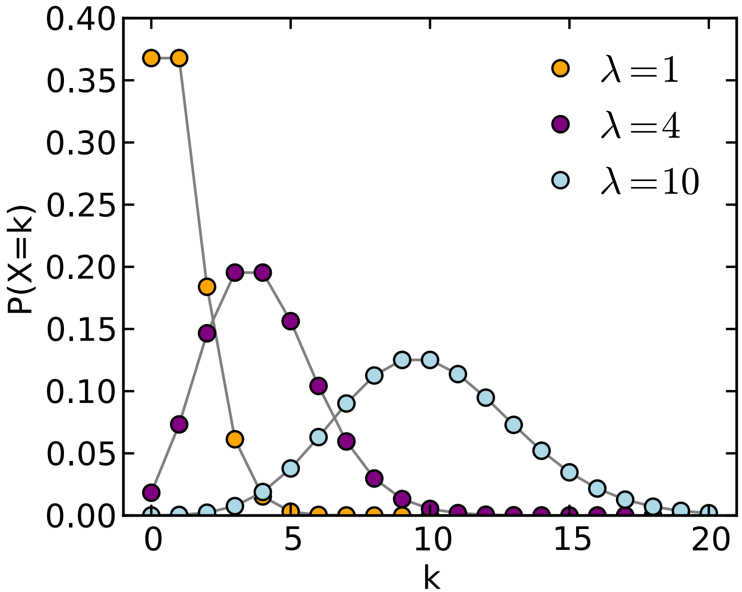
Probability

Poisson distribution

- A random variable X that takes on one of the values 0, 1, 2, . . . is said to be a Poisson random variable with parameter λ if, for some λ > 0,

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

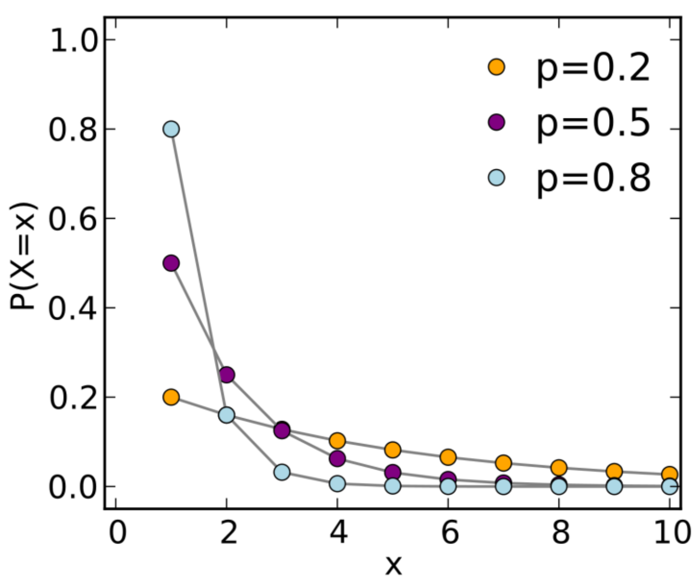
- $E(x)=\lambda$, $Var(x)=\lambda$
- Expresses the probability of a given number of events occurring in a fixed interval of time



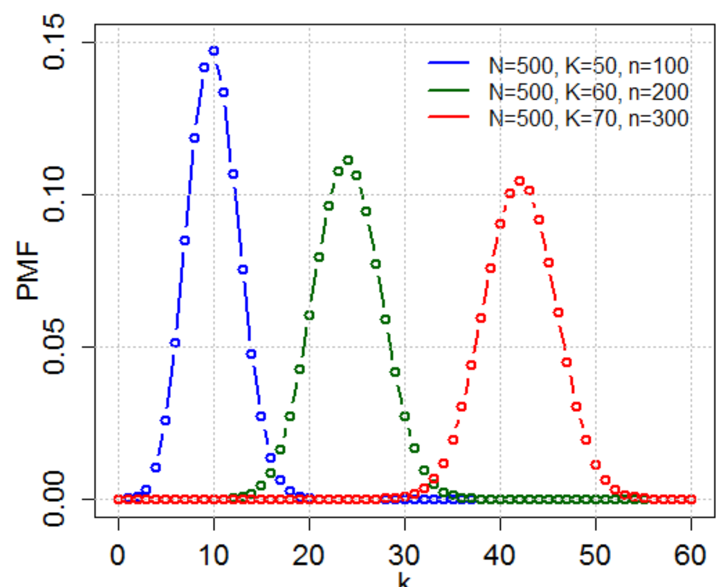
Probability

❑ Other discrete distributions

- Geometric distribution
- Hypergeometric distribution
- Negative binomial distribution



geometric distribution



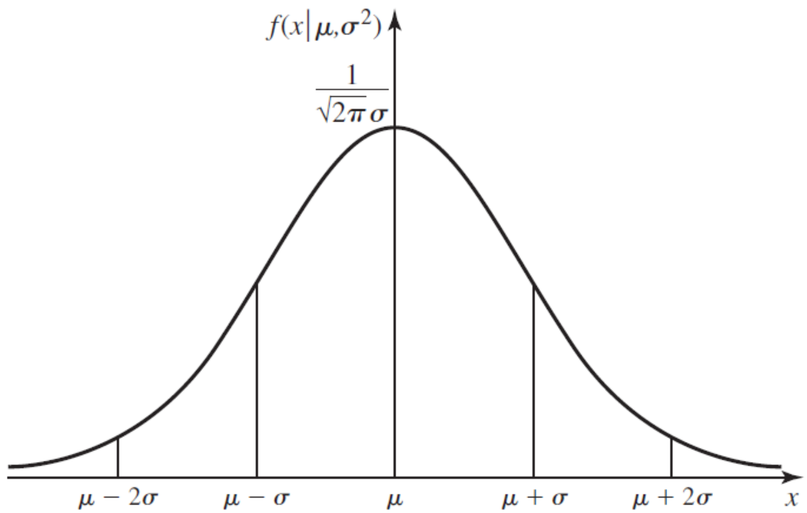
Hypergeometric distribution

Probability

Normal distribution

- A random variable *X* has the *normal distribution with mean μ and variance σ²* ($-\infty<\mu<\infty$ and $\sigma>0$) if *X* has a continuous distribution with the following p.d.f.:

$$f(x|\mu,\sigma^2)=\frac{1}{(2\pi)^{1/2}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\quad \text{for } -\infty<x<\infty.$$



Expectation

$$E(x)=\mu$$

Variance

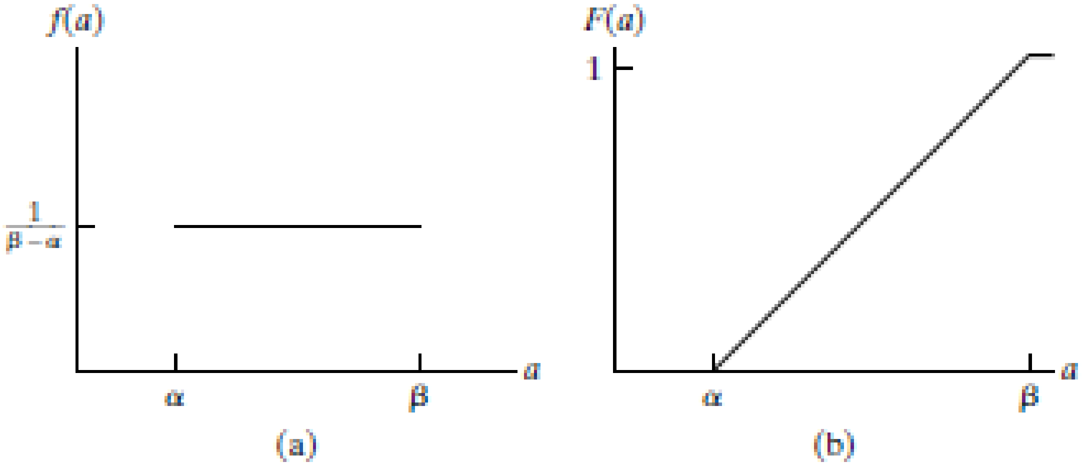
$$Var(x)=\sigma^2$$

Probability

❑ Uniform distribution

- X is a uniform random variable on the interval (α, β) if the probability density function of X is given by

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{if } \alpha < x < \beta \\ 0 & \text{otherwise} \end{cases}$$



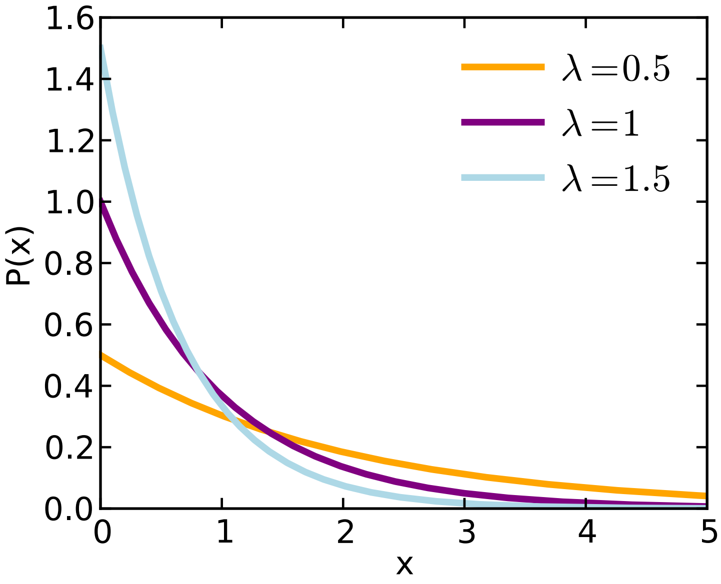
Graph of (a) $f(a)$ and (b) $F(a)$ for a uniform (α, β) random variable

Probability

Exponential distribution

- Let $\beta > 0$. A random variable *X* has the *exponential distribution with parameter β* if *X* has a continuous distribution with the p.d.f.

$$f(x|\beta) = \begin{cases} \beta e^{-\beta x} & \text{for } x > 0, \\ 0 & \text{for } x \le 0. \end{cases}$$



Expectation

$$E(x) = \frac{1}{\beta}$$

Variance

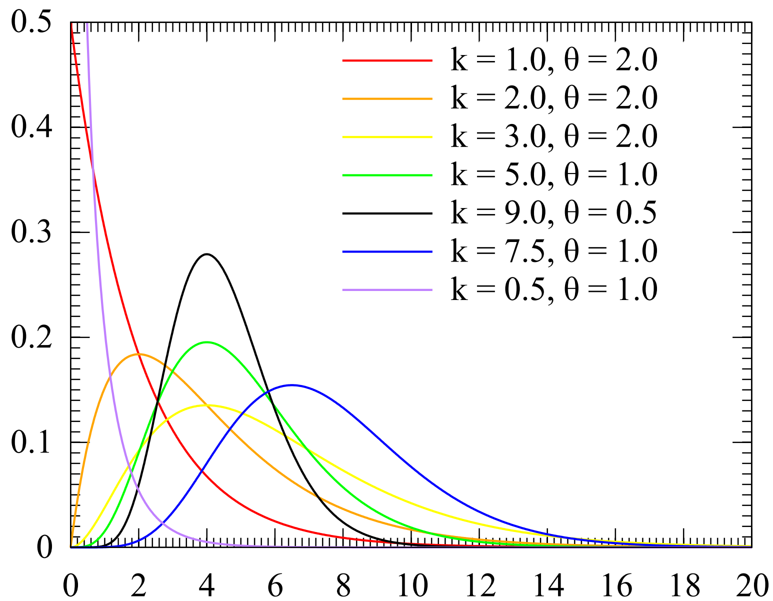
$$Var(x) = 1/\beta^2$$

Probability

Gamma distribution

- Let α and β be positive numbers.Arandom variable X has the *gamma distribution with parameters α and β* if X has a continuous distribution for which the p.d.f. is

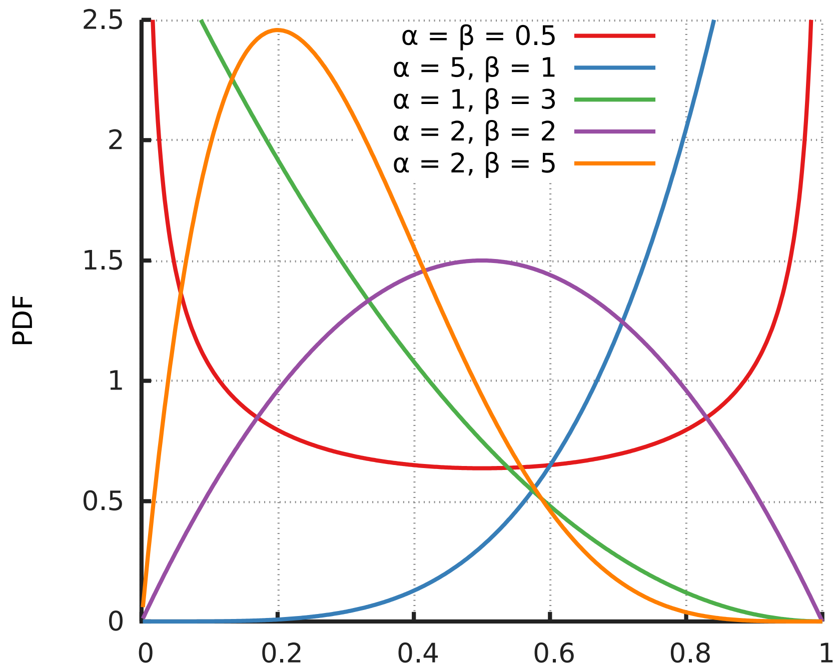
$$f(x|\alpha, \beta) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{for } x > 0, \\ 0 & \text{for } x \le 0. \end{cases}$$



Probability

- Beta distribution
  - Let α, $\beta > 0$ and let X be a random variable with p.d.f.

$$f(x|\alpha,\beta) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$



Probability

Joint Distribution

- Given random variables X,Y,..., that are defined on a probability space, the Joint distribution for X,Y,... is a probability that each of X,Y,...falls in any particular range or discrete set of values specified for that variable.
- Bivariate distribution vs multivariate distribution

[EXAMPLE]

Bivariate normal distribution

$$f(x_1, x_2) = \frac{1}{2\pi(1-\rho^2)^{1/2}\sigma_1\sigma_2} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}. \quad (5.10.2)$$

UNIVERSITY OF CALGARY

Probability

Law of Large Numbers

- Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the mean is μ and for which the variance is finite. Let X_n denote the sample mean. Then

$$\overline{X}_n \xrightarrow{p} \mu.$$

Central Limit Theorem

- If the random variables $X_1, \ldots, X_n$ form a random sample of size n from a given distribution with mean μ and variance $\sigma_2 (0 < \sigma_2 < \infty)$, then for each fixed number x,

$$\lim_{n\to\infty} \Pr\left[\frac{\overline{X}_n - \mu}{\sigma / n^{1/2}} \le x\right] = \Phi(x),$$

UNIVERSITY OF CALGARY

PART3 Introduction to Statistics

Statistics

- Contents of Statistics
  - Descriptive statistics
  - Statistical Inference
  - Point Estimation
  - Interval Estimation
  - Hypothesis Testing
  - Goodness of Fit
  - Analysis of Variance
  - Regression

Statistics

Statistic

- Let $X_1,\ldots X_n$ be a random sample of size n from a population and let $T(x_1,\ldots x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1,\ldots,X_n)$,the random variable or vector $Y=T(X_1,\ldots X_n)$ is called a Statistic.

Sampling distribution

- The probability distribution of a statistic Y is called the sampling distribution of Y

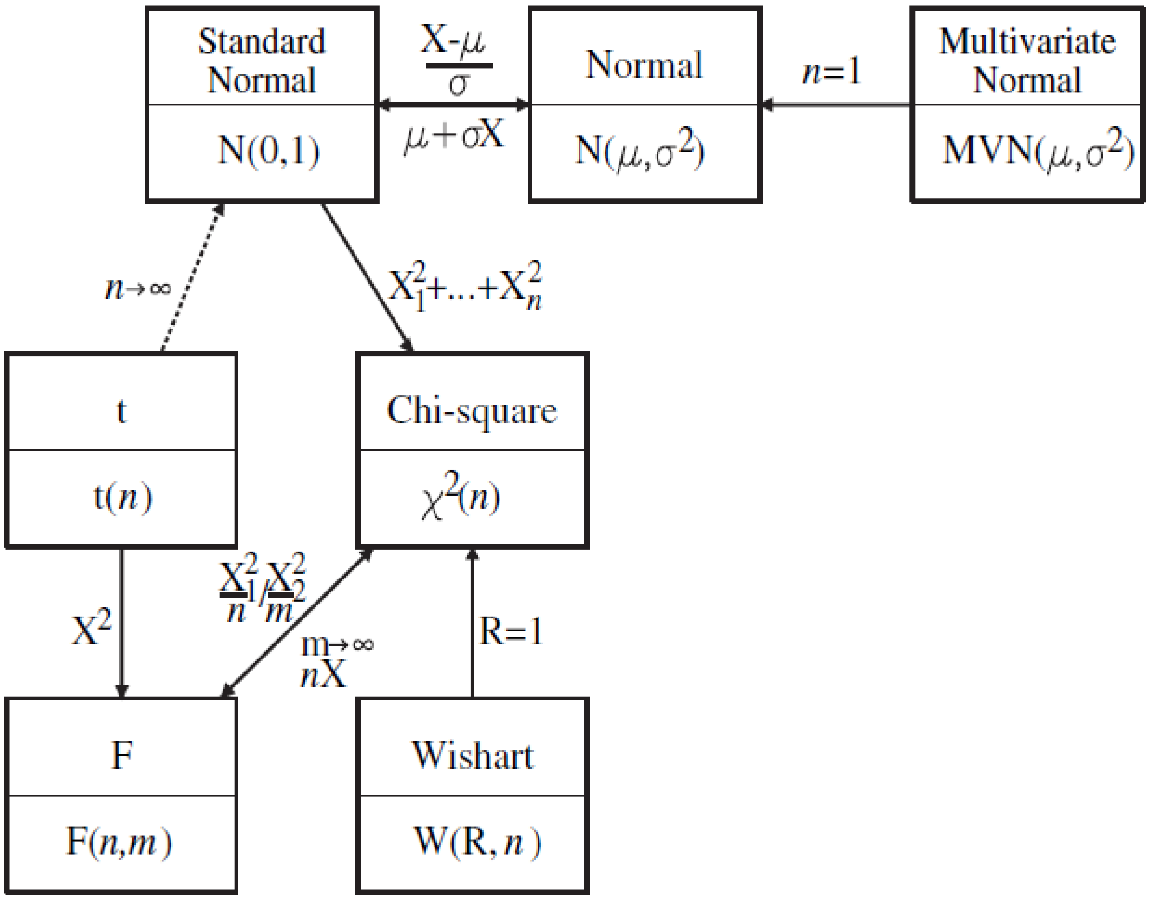
[EXAMPLE]

sample mean $\bar{X}$ and Sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ from samples of size n are two Statistics, If the population follows normal distribution, the sampling distribution of $\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$,$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, df = n-1$

Important sampling distribution: F distribution, Chi-square distribution and t distribution

UNIVERSITY OF CALGARY

Statistics



Standard Normal
N(0,1)
X-μ / σ
μ+σX
Normal
N(μ,σ2)
n=1
Multivariate Normal
MVN(μ,σ2)
n→∞
X12+...+Xn2
t
t(n)
Chi-square
χ2(n)
X2
X12/n / X22/m
m→∞
nX
R=1
F
F(n,m)
Wishart
W(R, n)

Statistics

- Statistical Inference
  - Parameter Estimation
    - Point estimation
      - method of moments
      - Maximum likelihood estimation
      - Bayesian estimation
    - Interval estimation
  - Hypothesis Test
  - Goodness of fit
  - ANOVA and Regression

Point Estimates

Method of moments

9.3 Definition. *The* method of moments estimator $\widehat{\theta}_n$ *is defined to be the value of* θ *such that*

$$\begin{aligned} \alpha_1(\widehat{\theta}_n) &= \widehat{\alpha}_1 \\ \alpha_2(\widehat{\theta}_n) &= \widehat{\alpha}_2 \\ \vdots \quad &\vdots \quad \vdots \\ \alpha_k(\widehat{\theta}_n) &= \widehat{\alpha}_k. \end{aligned} \tag{9.4}$$

[EXAMPLE]

9.4 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Then $\alpha_1 = \mathbb{E}_p(X) = p$ and $\widehat{\alpha}_1 = n^{-1} \sum_{i=1}^n X_i$. By equating these we get the estimator

$$\widehat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i. \quad \blacksquare$$

Point Estimates

Maximum Likelihood Estimation

设 $X=(X_1,\cdots,X_n)$ 为从具有概率函数 $f_\theta(x)$ 的总体中抽取的样本，θ 为未知参数或者参数向量. $x=(x_1,\cdots,x_n)$ 为样本的观察值。若在给定 x 时，值 $\hat{\theta}=\hat{\theta}(x)$ 满足下式

$$L(\hat{\theta})=\max_{\theta\in\Theta} L(x;\theta)$$

Definition

则称 $\hat{\theta}$ 为参数 θ 的最大似然估计值，而 $\hat{\theta}(X)$ 称为参数 θ 的最大似然估计量。若待估参数为 θ 的函数 $g(\theta)$，则 $g(\theta)$ 的最大似然估计量为 $g(\hat{\theta})$。

$$L(x;\theta)=\prod_{i=1}^{n} f_\theta(x_i) \longrightarrow l(\theta)=\log L(\theta) \longrightarrow \frac{\partial l(\theta)}{\partial \theta_i}=0$$

UNIVERSITY OF CALGARY

Statistics

❑ Point Estimates

❑ Maximum Likelihood Estimation

[EXAMPLE] Sampling from a Bernoulli Distribution

Likelihood function $$f_n(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}.$$

Loglikelihood function $$L(\theta) = \log f_n(\boldsymbol{x}|\theta) = \sum_{i=1}^{n}[x_i \log\theta + (1-x_i)\log(1-\theta)]$$

$$= \left(\sum_{i=1}^{n} x_i\right)\log\theta + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-\theta).$$

$$\frac{dL(\theta)}{d\theta} = 0 \longrightarrow \hat{\theta} = \bar{X}$$

Hypothesis test

事先对总体参数作出某种假设，然后利用样本信息来判断假设是否成立，有参数假设检验和非参数假设检验，均采用逻辑上的反证法，依据统计上的小概率原理

suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H0 : \theta \in \Theta_0 \text{ versus } H1 : \theta \in \Theta_1.$$

We call H_0 the null hypothesis and H_1 the alternative hypothesis

- 将受保护的对象置为零假设
- 如果你希望“证明” 某个命题, 就取相反结论或者其中一部分作为零假设

Statistics

Hypothesis test

决策＼事实	H_0 成立	H_1 成立
不拒绝 H_0	不犯错	第 II 类错误
拒绝 H_0	第 I 类错误	不犯错

因为犯第I类错误危害更大, 需要尽量避免犯第I类错误. 因此，这种在只限制第一类错误的原则下的检验方法，就称为“显著性检验“

具体地, 给定一个允许的犯第一类错误概率的最大值α, 选取τ使得

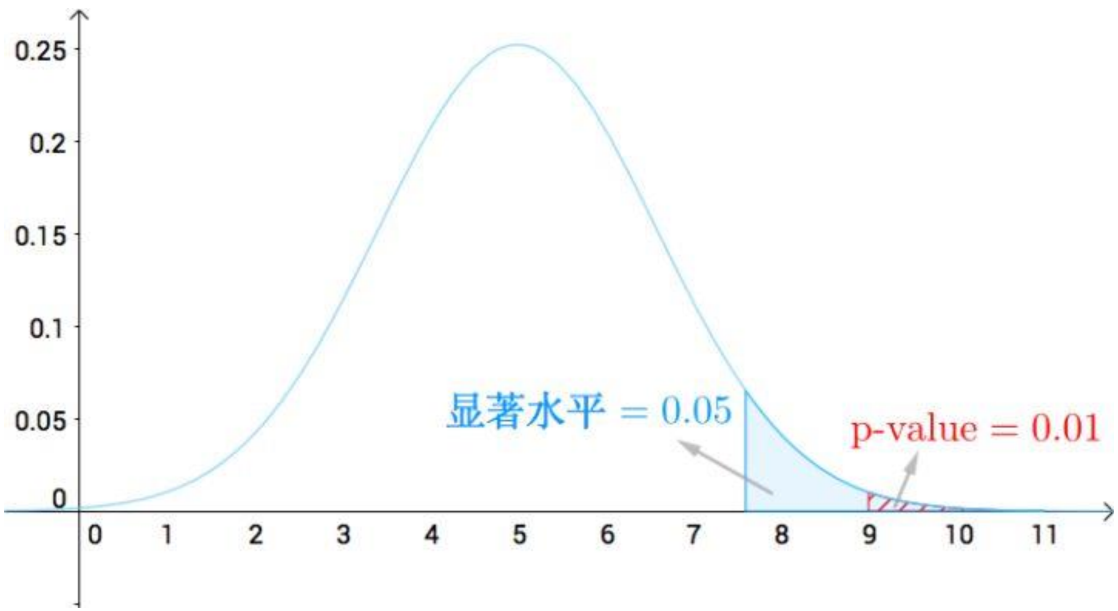
$$P_{H0}(T < \tau) \leq \alpha$$

称α为显著性水平（significant level）. 通常将 取为0.1, 0.05, 0.01等较小的数

UNIVERSITY OF CALGARY

Hypothesis test

p-value. In general, the *p-value* is the smallest level α_0 such that we would reject the null-hypothesis at level α_0 with the observed data.



Hypothesis test steps

1. 根据问题, 提出假设检验问题

$$H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1.$$

   其中 H_0 为**零假设**或**原假设**, 而 H_1 为**对立假设**或**备择假设**.

2. 根据参数的估计方法构造一个适当的**检验统计量** $T = T(X_1, \cdots, X_n)$, 其中 $X_1, \cdots, X_n$ 为从总体中抽得的一个样本.

3. 根据对立假设的形状构造一个检验的**拒绝域** $W = \{T(X_1, \cdots, X_n) \in A\}$, 其中 A 为一个集合, 通常是一个区间. 比如拒绝域可取为 $\{T(X_1, \cdots, X_n) > \tau\}$, 则称 τ 为 **临界值**.

4. 对任意的 $\theta \in \Theta_0$, 犯第 I 类错误的概率 $P_\theta(T(X_1, \cdots, X_n) \in A)$ 小于或等于某个指定正的常数 α), 则称 α 为**显著性水平**.

5. 结合 T 在 H_0 下的分布, 定出 A.

Statistics

- Goodness of fit

The goodness of fit of a statistical model describes how well it fits a set of observations.

Pearson's chi-squared test uses a measure of goodness of fit which is the sum of differences between observed and expected outcome frequencies (that is, counts of observations), each squared and divided by the expectation:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i = an observed frequency (i.e. count) for bin I

E_i = an expected (theoretical) frequency for bin i, asserted by the null hypothesis.

Likelihood ratio test and Wald test

UNIVERSITY OF CALGARY

PART4 An example

A generalized beta binomial distribution and example

Jiang Li and Yu Wang

December 7, 2018

UNIVERSITY OF CALGARY

OUTLINE

01

Problem with Beta Binomial Distribution

02

Generalized Beta Binomial Distribution

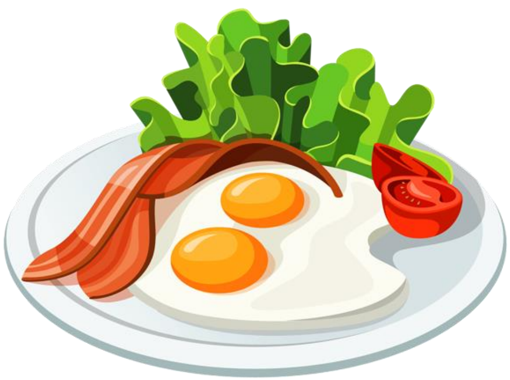
03

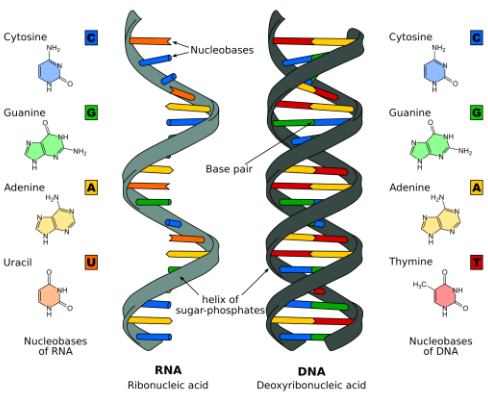
Alcohol Drinking days Sample

PART1 Problem with Beta Binomial Distribution

Beta Binomial Distribution

- The beta-binomial distribution is the binomial distribution in which the probability of success at each trial is fixed but randomly drawn from a beta distribution prior to *n* Bernoulli trials.
- Application area





Purchasing and consumption behavior

Alanko and Lemmens, 1996
Danaher and Hardie, 2005
Chatfield and Goodhart,1970

Spatial heterogeneity

Yang et al., 2005
Shiyomi et al., 2009

Bioinformatic research

RNA-seq experiments(C.A. Kapourani,2018)
Methylation analysis(Dolzhenko et al., 2014)

UNIVERSITY OF CALGARY

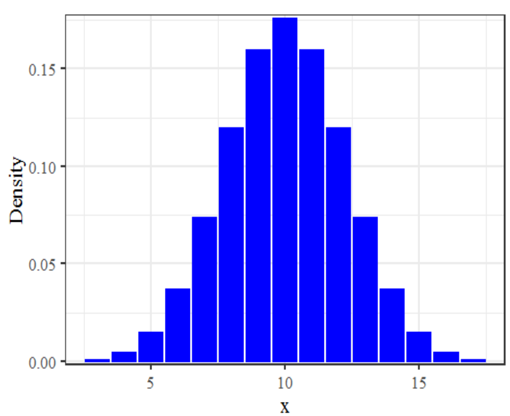
Beta Binomial Distribution

- ❑BB distribution fit better in count datasets with overdispersion
- ❑Beta distribution is the conjugate prior of BB distribution
- ❑Expanded the choice of models

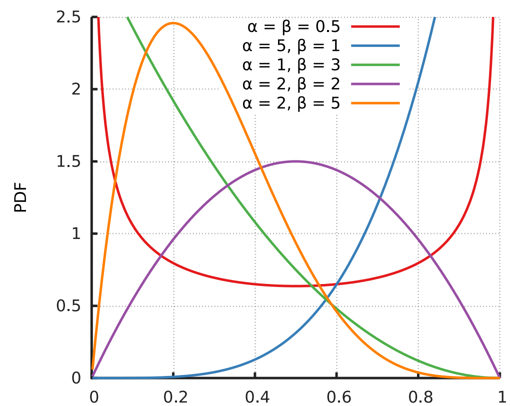
$$f_B(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$f_B(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

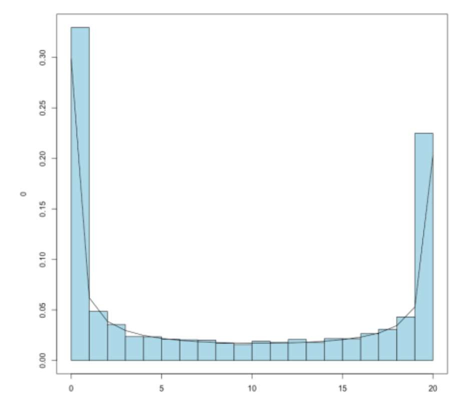
$$f_{BB}(x;\alpha,\beta) = \binom{n}{x} \frac{B(\alpha+x, n+\beta-x)}{\mathrm{B}(\alpha,\beta)}$$



Binomial distribution
X~Binomal(20,0.5)



Beta distribution

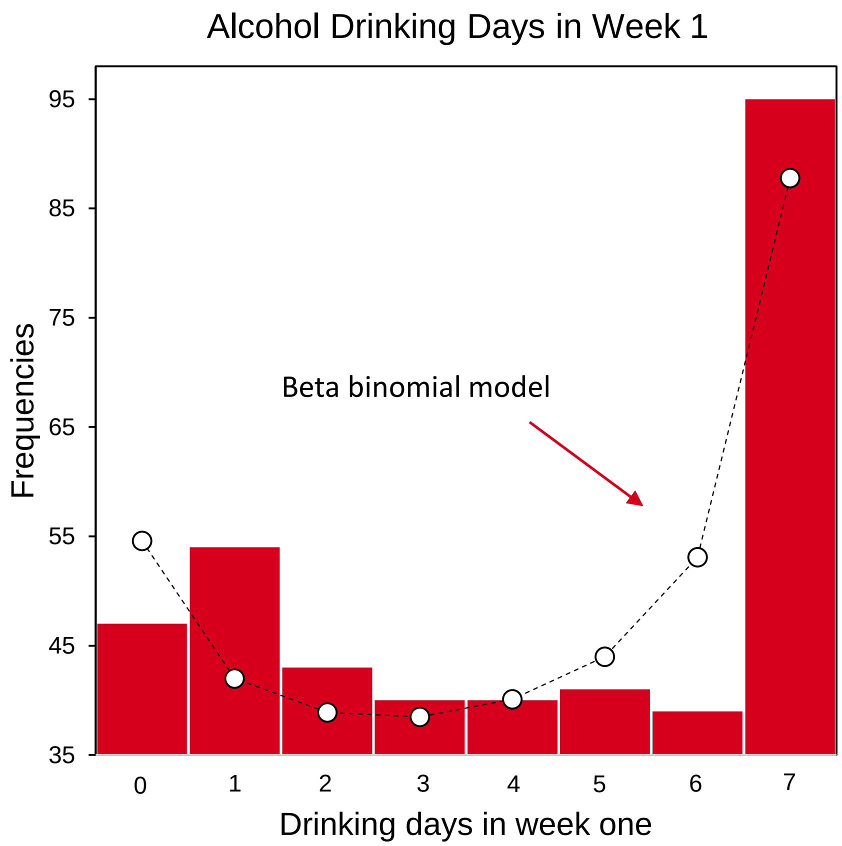


Beta binomial distribution
X~Beta binomial(20,0.2,0.25)

UNIVERSITY OF CALGARY

Beta Binomial Distribution

❑ Difficulties with Beta Binomial Models



- ❑ We need a new distribution
  - Fit better
  - Flexible
  - More parameters
  - Based on Beta binomial distribution

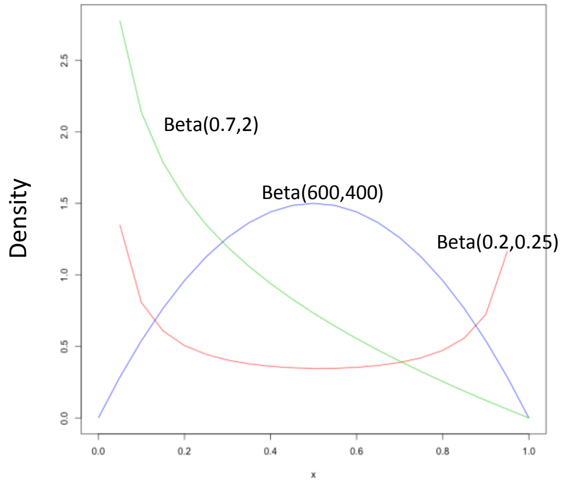
PART2 Generalized Beta Binomial Distribution

Beta Binomial Distribution

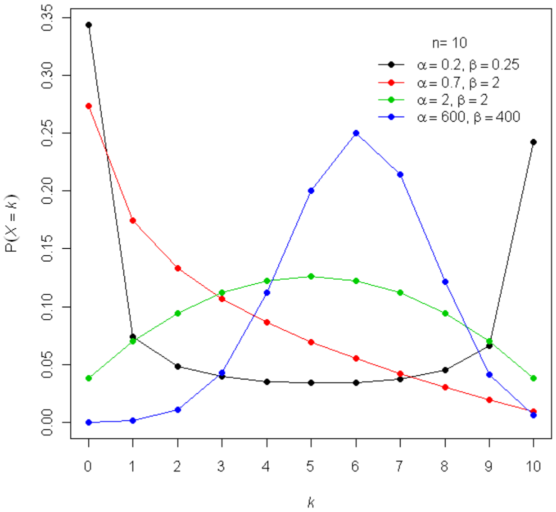
- Beta Binomial distribution is a beta mixture of the binomial distribution

$$Binomial(n,p) \underset{p}{\wedge} Beta(\alpha,\beta)$$

(Johnson et al., 1992)



Probability density function of Beta Distribution



Probability mass function of Beta Binomial Distribution

UNIVERSITY OF CALGARY

Generalized Beta Binomial Distribution

- Generalized Beta Distribution

$$f_P(p) = \frac{1}{B(\alpha,\beta)} \frac{{}_2F_1\,(-n,\alpha,-\beta-n+1;1)}{{}_2F_1\,(-n,\alpha,-\beta-n+1;\lambda)} p^{\alpha-1}(1-p)^{\alpha-1}\left[\frac{\lambda^{\beta+n}}{\{\lambda+(1-\lambda)p\}^{\alpha+\beta+n}}\right]$$

With $0<p<1$ and $\alpha,\beta,\lambda>0$

Where ${}_2F_1$ is the Gaussian hypergeometric function

- Generalized Beta Binomial Distribution

$$Binomial(n,p) \underset{p}{\wedge} \text{Generalized}Beta(\alpha,\beta,\lambda)$$

$$f_{GBB}(x;\alpha,\beta,\lambda) = f_0 \binom{n}{x} \frac{B(\alpha+x,n+\beta-x)}{\mathrm{B}(\alpha,\beta)} \lambda^x$$

where $$f_0 = \frac{1}{{}_2F_1\,(-n,\alpha;-\beta-n+1;\lambda)}$$

UNIVERSITY OF CALGARY

Generalized Beta Binomial Distribution

- The PMF of the BB distribution and its generalization only differ by a scale factor and the factor λ^x.

- While $\lambda=1$, Generalized beta binomial distribution degraded to beta binomial distribution

$$f(x) = \binom{n}{x} \frac{B(\alpha + x, n + \beta - x)}{B(\alpha, \beta)} \quad \text{vs} \quad f(x) = f_0 \binom{n}{x} \frac{B(\alpha + x, n + \beta - x)}{B(\alpha, \beta)} \lambda^x$$

$$f_0 = \frac{1}{{}_2F_1(-n, \alpha; -\beta - n + 1; \lambda)}$$

UNIVERSITY OF CALGARY

Generalized Beta Binomial Distribution

- Parameter λ extended the beta binomial distribution
- GBB concentrates probability in high values of the variable if λ > 1 or in low values if λ < 1

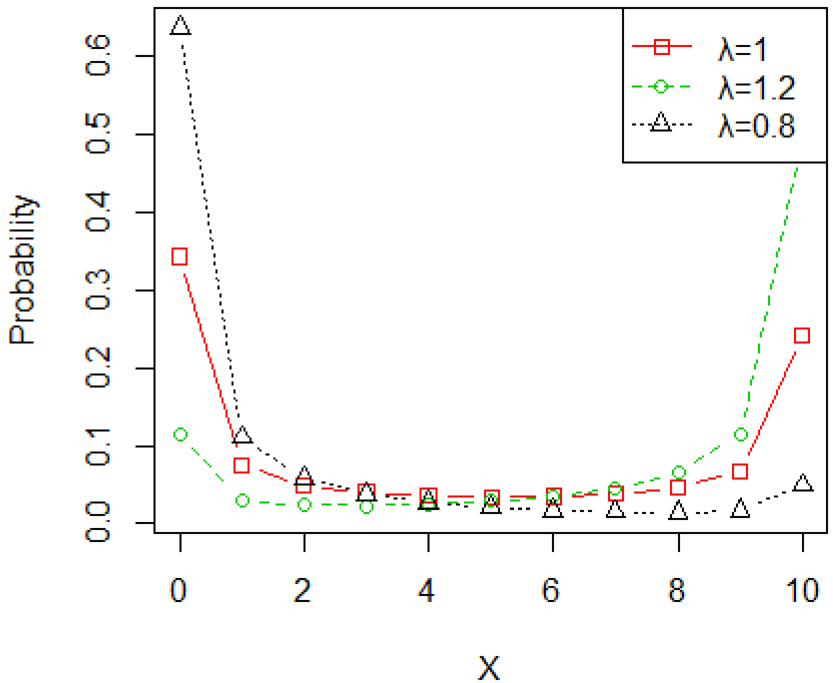


Figure: PMF of $GBB_{10}(0.2,0.25,0.8)$, $BB_{10}(0.2,0.25)$ and $GBB_{10}(0.2,0.25,1.2)$

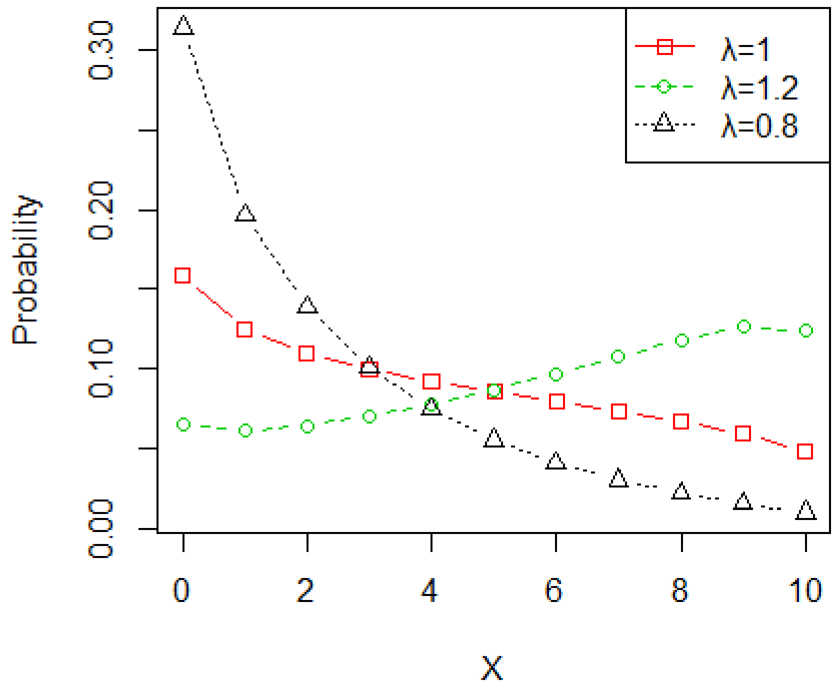


Figure: PMF of $GBB_{10}(0.8,1.2,0.8)$, $BB_{10}(0.8,1.2)$ and $GBB_{10}(0.8,1.2,1.2)$

UNIVERSITY OF CALGARY

PART3 Alcohol Drinking days Sample

Alcohol drinking days data

- Description of dataset:

  - The daily alcohol consumption data were collected from general population surveys conducted in Netherlands in 1983, 399 respondents was asked to keep a diary for consecutive days.

Table 1. Number of alcohol drinking days

Number of days per week	0	1	2	3	4	5	6	7
Observed frequencies in week one	47	54	43	40	40	41	39	95
Observed frequencies in week two	42	47	54	40	49	40	43	84

  - Significance

    the alcohol drinking days model can help retailers to analyze the alcohol consumption behavior, help health care provider to better predict the relationship between alcohol consumption and disease.

  - Methods: Beta Binomial Distribution and Generalized Beta Binomial Distribution

UNIVERSITY OF CALGARY

Model fitting

- MLE is used to derive the estimator of α,β,λ
  - Log likelihood function

$$\log\{L(\alpha,\beta,\lambda)\} = \sum_{x_i}\left[log\binom{n}{x_i} + \log\{beta(\alpha + xi, \beta + n - xi)\} + xi\log(\lambda)\right] - N\log\{beta(\alpha,\beta + n)\} + N\log(f_0)$$

```
log.lik<-function(p,x)
{
  library(hypergeo)
  log.lik=-(-length(x)*log(hypergeo(-7,p[1],-6-p[2]
,p[3]))+sum(log(beta(p[1]+x,p[2]+7-x)))-length(x)*l
og(beta(p[1],7+p[2]))+log(p[3])*sum(x))
}
p<-c(0.722,0.581,0.3)
optim(p, fn=log.lik,method="L-BFGS-B",x=x1)

## $par
## [1] 1.3506945 0.3245518 0.7005232
```

  - Estimates of parameters

Parameters	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\lambda}$	Log-likelihood
This Project	1.35069	0.32455	0.70052	-809.2767
Article	1.3508	0.3245	0.7005	-809.2767

UNIVERSITY OF CALGARY

Model fitting

- Estimates of Frequencies of alcohol drinking days

  - Probability of drinking days

$$P(X = xi) = \binom{n}{x_i} \frac{B(\hat{\alpha} + x, n + \hat{\beta} - xi)}{{}_2F_1(-n, \hat{\alpha}; -\hat{\beta} - n + 1; \lambda) B(\hat{\alpha}, \hat{\beta})} \hat{\lambda}^{x_i}$$

Where x_i =0,1,2,...,7

  - Estimates of frequency

$$\text{Estimated frequency} = n * P(X = xi)$$

```
#Estimates of week 1 frequencies
library(hypergeo)
coeff1=1/hypergeo(-7,1.3506945,-0.3245518-6,0.
7005232)
fact1=factorial(7)/(factorial(freq)*factorial(
7-freq))
expo1=(0.7005232)^freq
betanum1=beta(1.3506945+freq,0.3245518+7-freq)
ptms1=coeff1*fact1*betanum1*expo1/beta(1.35069
45,0.3245518+7)
est1=length(x1)*ptms1
## [1] 47.87456+0i 50.13635+0i 46.51678+0i 42.
07990+0i 38.57649+0i 37.32221+0i
## [7] 41.78505+0i 94.70867+0i
```

Numbers of drinking days		0	1	2	3	4	5	6	7
Project	Week 1	47.87	50.14	46.52	42.08	38.58	37.32	41.79	94.71
	Week 2	41.18	49.90	49.55	46.32	42.90	41.12	44.31	83.71
Article	Week 1	47.87	50.13	46.52	42.08	38.58	37.32	41.79	94.78
	Week 2	41.17	49.89	49.55	46.32	42.90	41.12	44.31	93.74

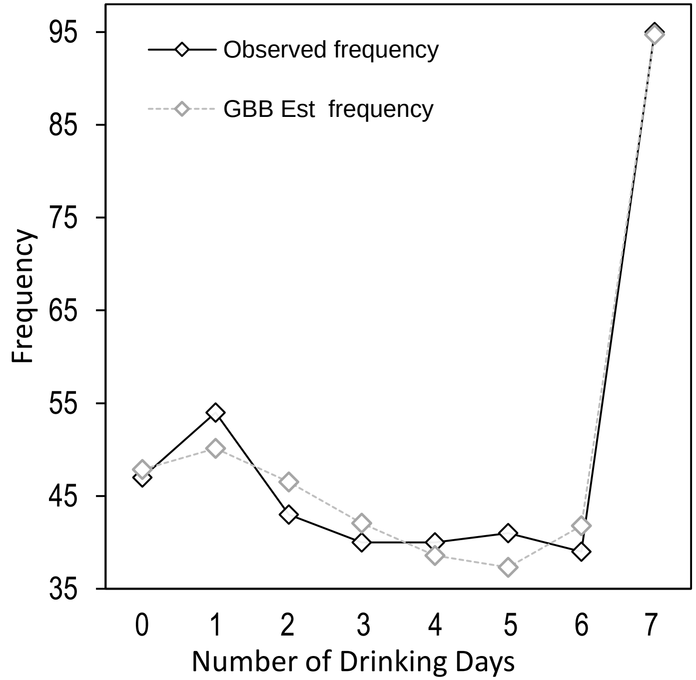
UNIVERSITY OF CALGARY

Goodness of fit

- Chi-square test

χ^2 statistics for week 1: $\chi^2 = \sum_{i=0}^{7} \frac{(O_i - E_i)^2}{E_i} = 1.2839 \ \ (df = 8 - 1 - 3 = 4)$

χ^2 statistics for week 2: $\chi^2 = \sum_{i=0}^{7} \frac{(O_i - E_i)^2}{E_i} = 2.3821 \ \ (df = 8 - 1 - 3 = 4)$



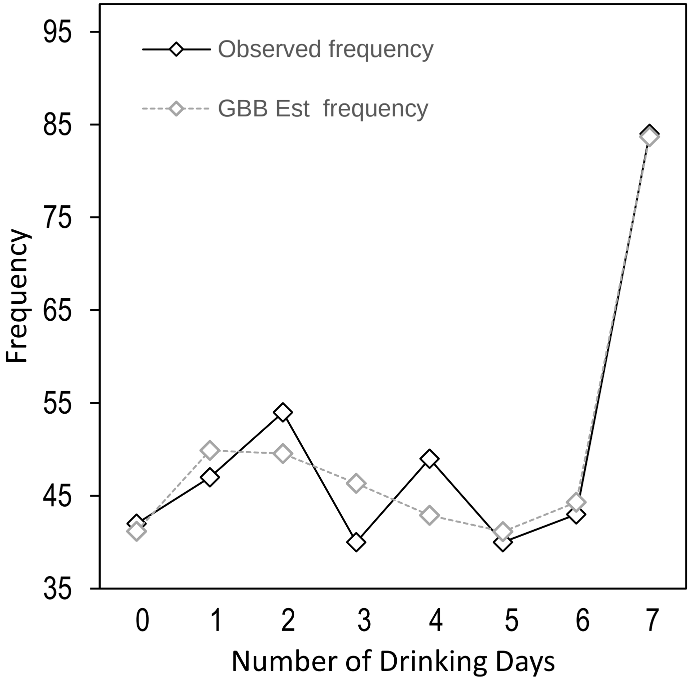


Figure: Observed frequency and fitted frequency by the GBB model

UNIVERSITY OF CALGARY

Comparing BB and GBB model

- Wald statistic

To test the hypothesis that the Beta Binomial Model is adequate (λ=1)

$$H_0: \lambda = 1 \ \ v.s \ \ H_1: \lambda \neq 1$$

Wald statistics for week 1: $T_w = \frac{(\widehat{\lambda_1}-1)^2}{Var(\widehat{\lambda_1})} = 10.1086$

Wald statistics for week 2: $T_w = \frac{(\widehat{\lambda_2}-1)^2}{Var(\widehat{\lambda_2})} = 42.2308$

The Wald test statistic is very significant when compared with $\chi^2(1)$. We reject the null hypothesis, so the BB model is inadequate for this dataset

Comparing BB and GBB model

- Likelihood ratio test

likelihood ratio test is used to contrast the BB distribution against GBB distribution,
null hypothesis assumes BB distribution is adequate in this case

$$H_0: BB\ is\ adequate\ v.s\ \ H_1: BB\ is\ inadequate$$

LR test statistics for week 1: $T_{LR} = 2 \times (l_{GBB} - l_{BB}) = 8.3608$

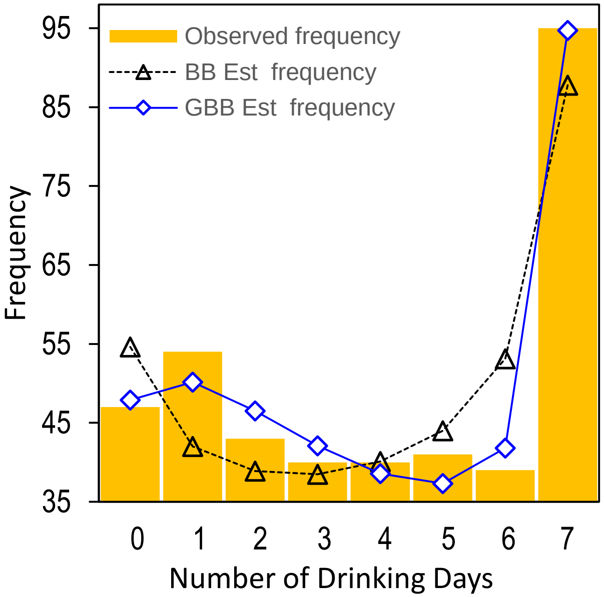
LR test statistics for week 2: $T_{LR} = 2 \times (l_{GBB} - l_{BB}) = 7.3626$

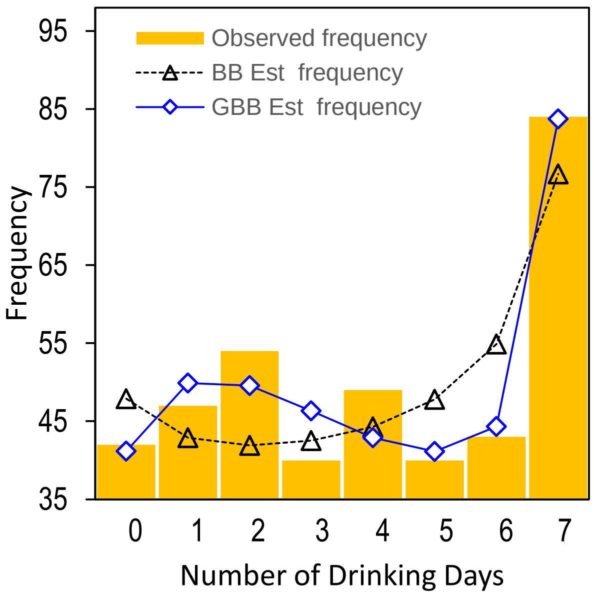
T_{LR} is very significant when compared with $\chi^2(1)$. Thus, we reject null hypothesis. Therefore, GBB is more appropriate for fitting this kind of data

UNIVERSITY OF CALGARY

Summary

- The fits that were obtained by means of the GBB distribution are a substantial improvement on those obtained by the BB distribution
- Pearson, Wald and Likelihood Ratio test also gives the same conclusion
- Estimated λ is less than 1, probability concentrates in low values which the variable could take
- The parameter λ could be used to detect overdispersion of data, and GBB could better fit some overdispersion data





$$f(x) = f_0 \binom{n}{x} \frac{B(\alpha + x, n + \beta - x)}{B(\alpha, \beta)} \lambda^x$$

UNIVERSITY OF CALGARY

References Books

- Introduction to Probability, Grinstead and Snell
- Probability, Jim Pitman
- All of Statistics A Concise Course in Statistical Inference, Larry Wasserman
- Probability and Statistics, Morris H. DeGroot, Mark J. Schervish