Statistical Inference I and II

Probabilistic and Statistical Models

Stanislav Katina[1]

[1]Institute of Mathematics and Statistics, Masaryk University
Honorary Research Fellow, The University of Glasgow

November 13, 2018

Probabilistic and Statistical Models

Random variable, random vector, data, individuals

- **random variable** and **random vector**
  - *random variable* X is a function from a **sample space** to a set of real numbers $X : \mathcal{Y} \to \mathbb{R}$ (a set of all possible outcomes)
  - 2-dimensional *random vector* $(X_1, X_2)^T : \mathcal{Y} \to \mathbb{R}^2$
  - k-dimensional *random vector* $(X_1, X_2, \ldots, X_k)^T : \mathcal{Y} \to \mathbb{R}^k$
- **data** – *data vector* and *data matrix* – the elements of a vector and the rows of a matrix are measured on ***individuals*** (***statistical units***)
  - *data* as realisations of X – n-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, where n is a *sample size*
  - *data* as realisations of $(X_1, X_2)^T$ – $(n \times 2)$-dimensional matrix with rows $(x_{i1}, x_{i2})^T$, $i = 1, 2, \ldots, n$ and columns $\mathbf{x}_1$ and $\mathbf{x}_2$
  - *data* as realisations of $(X_1, X_2, \ldots, X_k)^T$ – $(n \times k)$-dimensional matrix with rows $(x_{i1}, x_{i2}, \ldots, x_{ik})^T$, $i = 1, 2, \ldots, n$ and columns $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_k$

Probabilistic and Statistical Models

Model

- based on **probabilistic sampling principles**, the *individuals* are sampled from a *population*
- **attribute** – a specific value of a variable
- with certain precision, data are **measured** on individuals
- **descriptive statistics** – describing and summarising data
- **inferential statistics** (**statistical inference**) – inferring (drawing conclusions) about random variable ***based on a model fitted to data***
- $\mathcal{F}$ is a **set of models** (probabilistic or statistical)
  - X is characterised by a model $F(\cdot)$, $F \in \mathcal{F}$
  - $(X_1, X_2)^T$ is characterised by a model $F^{(2)}(\cdot)$, $F \in \mathcal{F}$
  - $(X_1, X_2, \ldots, X_k)^T$ is characterised by a model $F^{(k)}(\cdot)$, $F \in \mathcal{F}$
- **parameter** – a numerical quantity that characterises a model – one-dimensional parameter θ, k-dimensional vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^T$

Probabilistic and Statistical Models

Distribution function, probability and density function

- useful assumption – $X_i, i = 1, 2, \ldots, n$, are **independently identically distributed** random variables
- **distribution function**
  - discrete random variable

$$F_X(x) = \Pr(X \leq x) = \sum_{i:x_i \leq x} \Pr(X = x_i),$$

  where $\sum_{i=1}^{k(\infty)} p_i = 1$, $\Pr(X = x_i) = p_i = f_X(x_i) = f(x_i), \forall x_i$, where p_i is **probability mass function**; $\{x_i, p_i\}_{i=1}^{k(\infty)}$, $k \in \mathbb{N}^+$
  - continuous random variable

$$F_X(x) = \int_{-\infty}^{x} f(t)\, dt, f(x) \geq 0,$$

  where $\int_{-\infty}^{\infty} f(x)\, dx = 1$, $f_X(x) = f(x) = \frac{\partial}{\partial x} F_X(x)$ is **density function**

- Θ is a **parametric space**, the **support** of $F(\cdot;\boldsymbol{\theta})$ is $\mathcal{Y}_{\boldsymbol{\theta}} \subseteq \mathbb{R}^n$ (the smallest set, where the distribution function is defined); sample space $\mathcal{Y} = \cup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\mathcal{Y}_{\boldsymbol{\theta}}$
- $\mathcal{F}$ as **a parametric set of distribution functions**

$$\mathcal{F} = \left\{F(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k\right\},$$

- $\mathcal{F}$ as **a parametric set of probability or density functions**

$$\mathcal{F} = \left\{f(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k\right\}$$

- $\mathcal{F}$ as **non-parametric set**

$$\mathcal{F} = \{\text{a set of all density functions}\},$$

alternatively, probability or distribution function can be used

- the term "**probability model**" is often reduced to "**distribution**"
- "Random variable X is distributed as $F(x)$" or "random variable X is characterised by distribution $F(x)$", notation $X \sim F_X(x)$; symbol "$\sim$" means "asymptotically", "for sufficiently large n" (notation $X \sim f_X(x)$ is used very rarely)
- "Random variable X is distributed as random variable Y" or "Random variable X and Y are identically distributed" (notation $X \sim Y$ or $F_X(x) \sim F_Y(y)$)
- the term "**statistical model**" is often reduced to "**model**" (usually referred as ***causal statistical model*** or ***model of causal dependence***)
- "Y depends on X", where X is **independent variable** and Y is **dependent variable** (notation $Y|X$)

- "X is normally distributed with parameters μ and σ^2", notation $X \sim N(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma^2)^T$
- "$\mathbf{X} = (X_1, X_2)^T$ is characterised by bivariate normal distribution with parameters μ_1, μ_2, σ_1^2, σ_2^2 and ρ", notation $X \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$
- "$\mathbf{X} = (X_1, X_2, \ldots, X_k)^T$ is characterised by multivariate normal distribution with parameters μ_1, μ_2, ..., μ_k, σ_1^2, σ_2^2, ..., σ_k^2, and $\rho_{1,2}$, ..., $\rho_{k-1,k}$, ", notation $X \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta} = (\mu_1, \mu_2, \ldots, \mu_k, \sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2, \rho_{1,2}, \ldots, \rho_{k-1,k})^T$
- "X is binomially distributed with parameter p", notation $X \sim Bin(N, p)$, where $\theta = p$
- "X is characterised by distribution with parameter λ", notation $X \sim Poiss(\lambda)$, where $\theta = \lambda$
- "$\mathbf{X} = (X_1, X_2, \ldots, X_k)^T$ is multinomially distributed with parameter $\mathbf{p}$", notation $\mathbf{X} \sim Mult_k(N, \mathbf{p})$, where $\boldsymbol{\theta} = \mathbf{p}$

- "X is normally distributed with parameters μ and σ^2", notation $X \sim N(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma^2)^T$
- Random variable Z (Z-transformation) $\Pr(Z = \frac{X-\mu}{\sigma} < x_{1-\alpha}) = 1 - \alpha, Z \sim N(0, 1)$
- Rule "$90 - 95 - 99$" $\Pr(a \leq X \leq b) = 1 - \alpha$, where $1 - \alpha = 0.90, 0.95$ and 0.99, $a = \mu - x_{1-\frac{\alpha}{2}}\sigma$ and $b = \mu + x_{1-\frac{\alpha}{2}}\sigma$
- Rule "$68.27 - 95.45 - 99.73$" $\Pr(a \leq X < b) = \Pr(X < b) - \Pr(X < a) = F_X(b) - F_X(a)$, where $a = \mu - k\sigma$, $b = \mu + k\sigma$, $k = 1, 2$ and 3

Definition (approximation of binomial distribution by normal distribution)

If random variable X is binomially distributed with parameter p, $X \sim Bin(N,p)$, where $\theta = p$, if $Np > 5$ and $Nq > 5$, where $q = 1 - p$, then the distribution of random variable X can be approximated by normal distribution, $X \sim N(Np, Npq)$, where $\boldsymbol{\theta} = (Np, Npq)^T$.

Table: Examples of minimal N for fixed p

p	0.1	0.2	0.3	0.4	0.5
q	0.9	0.8	0.7	0.6	0.5
N	51	26	17	13	11

Definition (approximation of binomial distribution by normal distribution, Hald condition)

If random variable X is binomially distributed with parameter p, $X \sim Bin(N,p)$, where $\theta = p$, if $Npq > 9$ (Hald condition), where $q = 1 - p$, then the distribution of random variable X can be approximated by normal distribution, $X \sim N(Np, Npq)$, where $\boldsymbol{\theta} = (Np, Npq)^T$.

Table: Examples of minimal N for fixed p

p	0.01	0.02	0.05	0.10	0.15	0.20	0.30	0.40	0.50
$1-p$	0.99	0.98	0.95	0.90	0.85	0.80	0.70	0.60	0.50
N	910	460	190	100	71	57	43	38	36

Example

Let $\Pr(\text{male}) = 0.515$ and $\Pr(\text{female}) = 0.485$. Let X be the frequency of males and Y frequency of females. Assuming that $X \sim Bin(N,p)$, calculate (a) $\Pr(X \leq 3)$, if $N = 5$, (b) $\Pr(X \leq 5)$, if $N = 10$ and (c) $\Pr(X \leq 25)$, if $N = 50$. Compare the results with normal approximation $X \sim N(Np, Npq)$.

Solution

(a) $E[X] = Np = 5 \times 0.515 = 2.575, E[Y] = 5 \times 0.485 = 2.425$,
$\Pr(X \leq 3) = \sum_{k \leq 3} \binom{5}{k} 0.515^k 0.485^{5-k} = 0.793$,
$\Pr(X \leq 3) = 0.648, N(5 \times 0.515, 5 \times 0.515 \times 0.485)$.
(b) $E[X] = 10 \times 0.515 = 5.15, E[Y] = 10 \times 0.485 = 4.85$,
$\Pr(X \leq 5) = \sum_{k \leq 5} \binom{10}{k} 0.515^k 0.485^{10-k} = 0.586$,
$\Pr(X \leq 5) = 0.462, N(10 \times 0.515, 10 \times 0.515 \times 0.485)$.
(c) $E[X] = 50 \times 0.515 = 25.75, E[Y] = 50 \times 0.485 = 24.25$,
$\Pr(X \leq 25) = \sum_{k \leq 25} \binom{50}{k} 0.515^k 0.485^{50-k} = 0.471$,
$\Pr(X \leq 25) = 0.416, N(50 \times 0.515, 50 \times 0.515 \times 0.485)$.

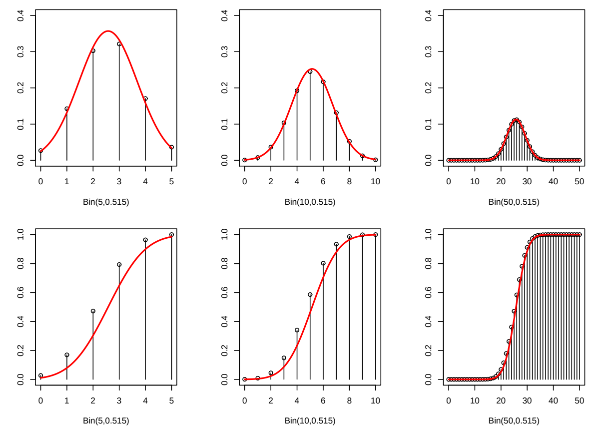


Figure: Probability function (first row) and distribution function (second row) of binomial distribution superimposed by normal distribution ($p = 0.515$; $N = 5, 10$ and 50)

Probabilistic and Statistical Models
Approximation of binomial distribution by normal distribution

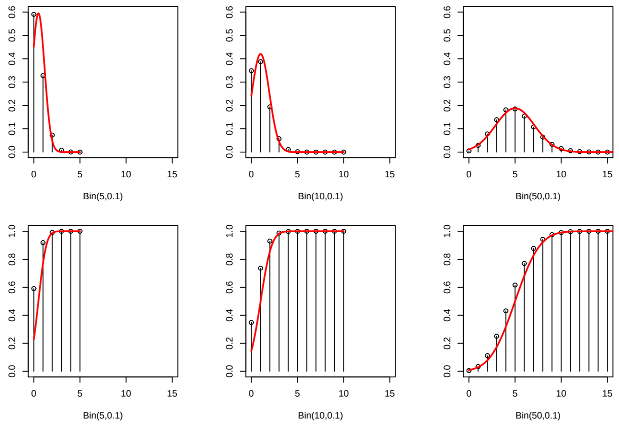


Figure: Probability function (first row) and distribution function (second row) of binomial distribution superimposed by normal distribution ($p = 0.1$; $N = 5, 10$ and 50)

Probabilistic and Statistical Models
Binomial distribution

Example (number of boys)

Number of boys X in families with N children is binomially distributed, i.e. $X \sim Bin(N, p)$, where $N = 12$, number of families $M = 6115$ (Geissler 1889). **Question**: Calculate theoretical frequencies $m_{n,E}$.

You know that $p = \frac{\sum_{n=0}^{N} n m_{n,O}}{NM} = 0.5192$ (weighted average; average of number of families weighted by number of boys).

Table: Observed and theoretical frequencies ($m_{n,O}$ and $m_{n,E}$) of families with n boys (O = observed, E = expected, theoretical)

n	0	1	2	3	4	5	6	7	8	9	10	11	12
$m_{n,O}$	3	24	104	286	670	1033	1343	1112	829	478	181	45	7
$m_{n,E}$	1	12	72	**259**	628	1085	1367	1266	854	410	133	26	2

Probabilistic and Statistical Models
Binomial distribution

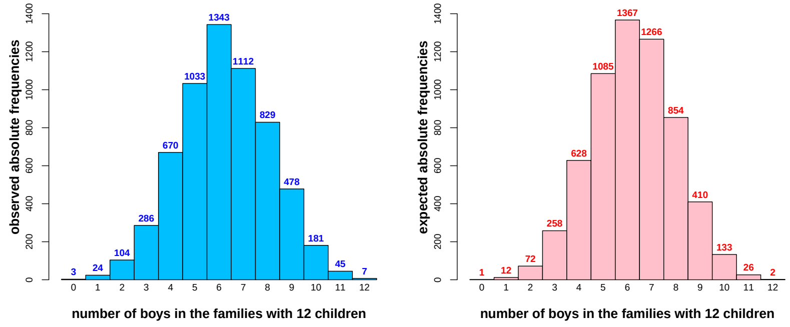


Figure: Histograms of observed and expected frequencies

Probabilistic and Statistical Models
Product-multinomial distribution

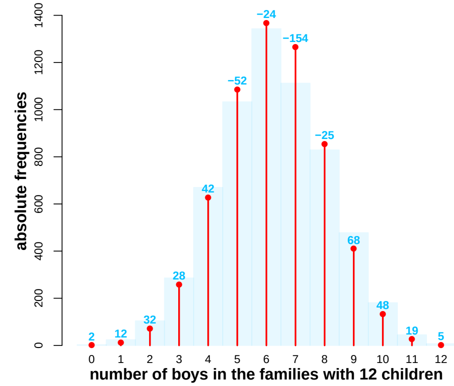


Figure: Comparison of observed and expected frequencies

Example (number of individuals with certain blood type)

Number of individuals $\mathbf{X} = (X_1, X_2, X_3, X_4)^T$ with certain blood group is multinomially distributed following Hardy-Wienberg equilibrium, i.e. $\mathbf{X} = (X_1, X_2, X_3, X_4)^T \sim Mult_4(N, \mathbf{p})$, where $N = 500$ (Katina et al. 2015). **Question**: Calculate theoretical frequencies $n_{j,E}$.

attributes (groups)	0	A	B	AB
$n_{j,O}$	209	184	81	26
$n_{j,E}$	210	183	80	27

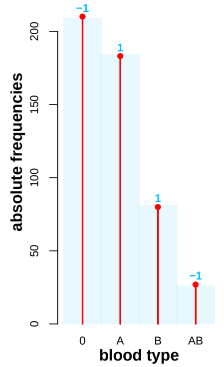


Figure: Comparison of observed and expected frequencies

Example (number of individuals with certain blood type)

Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^T$, where $\mathbf{X}_1 = (X_{11}, X_{12}, X_{13}, X_{14})^T$ is number of individuals in Košice (Slovakia) with certain blood group, $\mathbf{X}_2 = (X_{21}, X_{22}, X_{23}, X_{24})^T$ is number of individuals in Prague (Czech Republic) with certain blood group. $\mathbf{X}$ is product-multinomially distributed, i.e. $\mathbf{X} \sim ProdMult_2(\mathbf{N}, \mathbf{p})$, where $\mathbf{N} = (N_1, N_2)^T$, where $N_1 = 500$ and $N_2 = 400$ (Katina et al. 2015). Calculate theoretical frequencies $n_{E,ij}$. **Question**: What are the probabilities of having particular blood group in Prague and Košice?

Table: Observed frequencies of particular blood group

attributes (groups)	0	A	B	AB
$n_{1j,O}$=$n_{\text{Košice},j,O}$	138	147	84	31
$n_{2j,O}$=$n_{\text{Prague},j,O}$	209	184	81	26

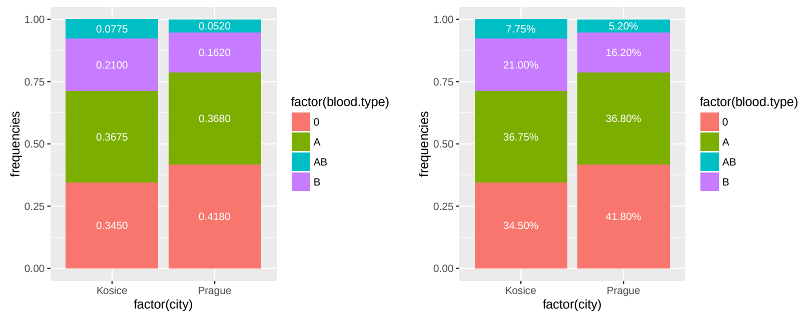


Figure: Barplots of four blood types in Košice and Prague (default palette)

Example (number of individuals with certain eye and hair colour)

Let $\mathbf{X} = (X_1, X_2, \ldots, X_{12})^T$ be random vector of number of individuals, eye colour (with levels blue Bl, green Gr, brown Br) and hair colour (with levels blond Blo, light-brown LB, black Bla, red R), where X_1 means Bl-Blo, X_2 means Bl-LB, X_3 means Bl-Bla, X_4 means Bl-R, X_5 means Gr-Blo, X_6 means Gr-LB, X_7 means Gr-Bla, X_8 means Gr-R, X_9 means Br-Blo, X_{10} means Br-LB, X_{11} means Br-Bla and X_{12} means Br-R. Let $\mathbf{X} \sim Mult_{12}(N, \mathbf{p})$, where $N = 6800$ (Yule and Kendall 1950).

Question: Calculate probabilities of having (1) particular eye and hair colour, (2) particular hair colour conditional on eye colour, (3) particular eye colour conditional on hair colour.

Table: 3×4 contingency table of frequencies n_j

	Blo	LB	Bla	R	row sums
Bl	1768	807	189	47	2811
Gr	946	1387	746	53	3132
Br	115	438	288	16	857
column sums	2829	2632	1223	116	6800

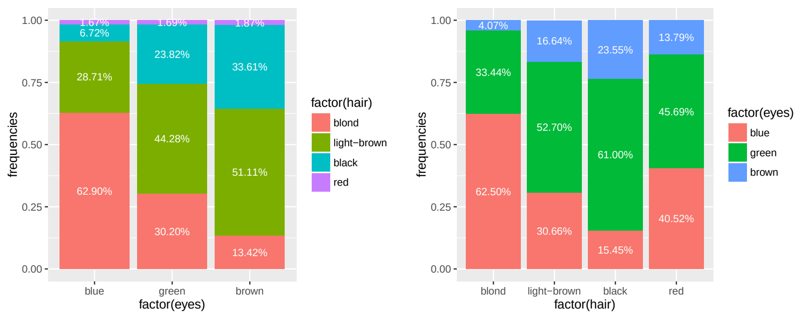


Figure: Barplots of eye and hair colour (default palette)

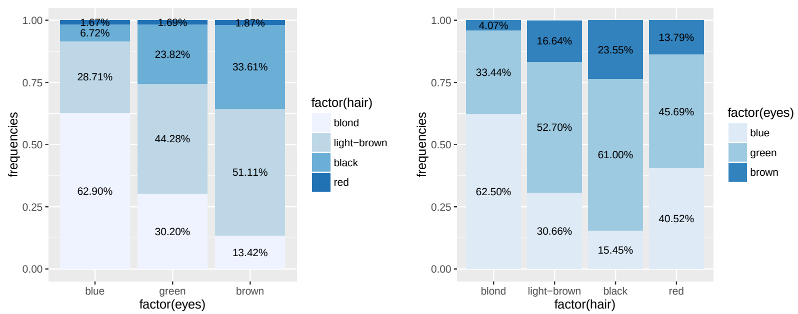


Figure: Barplots of eye and hair colour (blue palette)

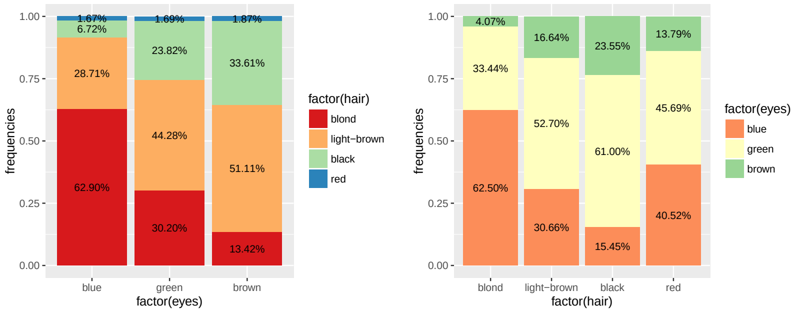


Figure: Barplots of eye and hair colour (spectral palette)

Example (number of individuals with certain socioeconomic status, political philosophy and political affiliation)

Number of individuals $X_1, \ldots, X_8$ with socioeconomic status, political philosophy and political affiliation is multinomially distributed, i.e. $\mathbf{X} = (X_1, \ldots, X_8)^T \sim Mult_8(N, \mathbf{p})$, where realisations $\mathbf{x} = (x_1, x_2, \ldots, x_8)^T$ and $N = 500$ (Christensen 1990, modified). **Question**: Calculate probabilities of having particular socioeconomic status, political philosophy and political affiliation.

Notation: (1) socioeconomic status (high – H, low – Lo), (2) political philosophy (democrat – D, republican – R) a (3) political affiliation (conservative – C, liberal – Li). Then X_1 (H-D-C), X_2 (H-D-Li), X_3 (H-R-C), X_4 (H-R-Li), X_5 (Lo-D-C), X_6 (Lo-D-Li), X_7 (Lo-R-C) and X_8 (Lo-R-Li).

Table: 2×4 contingency table of frequencies X_j

	D-C	D-Li	R-C	R-Li
H	60	60	60	20
Lo	90	90	90	30

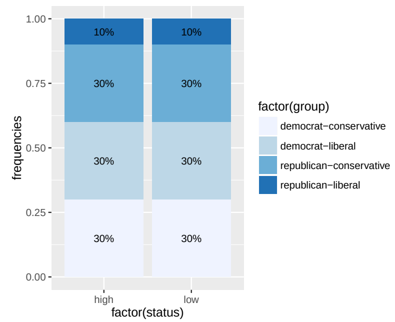


Figure: Barplots of socioeconomic status, political philosophy and affiliation (blue palette)

Example (Poisson distribution; killing by horse kicks)

Data were published by Russian economist *Ladislaus Bortkiewicz* in his book entitled *Das Gesetz der keinem Zahlen* (The Law of Small Numbers) in 1898. Let X be the number of corps of soldiers with n annual deaths (killed by horse kicks) in the Prussian army within one year (von Bortkiewicz 1898; in 10 different army corps; in 20 years, between 1875 and 1894), n be the number of annual deaths, $m_{n,O}$ be the number of army corps with particular number of annual deaths, $M = \sum_n m_{n,O} = 10 \times 20 = 200$. Then $X \sim Poiss(\lambda)$, where $\lambda = \frac{\sum_n nm_{n,O}}{\sum_n m_{n,O}} = 0.61$ (weighted average; average of number of army corps weighted by number of annual deaths). **Question**: Calculate theoretical frequencies $m_{n,E}$.

Table: Observed and theoretical frequencies ($m_{n,O}$ and $m_{n,E}$) of corps of solders with n annual deaths (killed by horse kicks) over 20 years

n	0	1	2	3	4	≥ 5
$m_{n,O}$	109	65	22	3	1	0
$m_{n,E}$	109	66	20	4	1	0

Probabilistic and Statistical Models
Poisson distribution

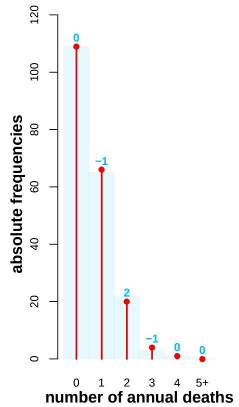


Figure: Comparison of observed and expected frequencies

Probabilistic and Statistical Models
Poisson distribution

Example (Poisson distribution; accidents in the factories)

Let X be the number of workers having an accident in munition factories in England during First World War (Greenwood and Yule 1920), n be the number of accidents, $m_{n,O}$ be the number of workers with particular number of accidents, $M = \sum_n m_{n,O} = 647$. Then $X \sim Poiss(\lambda)$, where $\lambda = \frac{\sum_n nm_{n,O}}{\sum_n m_{n,O}} = 0.47$ (weighted average; average of number of workers weighted by number of accidents). **Question**: Calculate theoretical frequencies $m_{n,E}$.

Table: Observed and theoretical frequencies ($m_{n,O}$ and $m_{n,E}$) of workers with n accidents

n	0	1	2	3	4	≥ 5
$m_{n,O}$	447	132	42	21	3	2
$m_{n,E}$	406	189	44	7	1	0

Probabilistic and Statistical Models
Poisson distribution

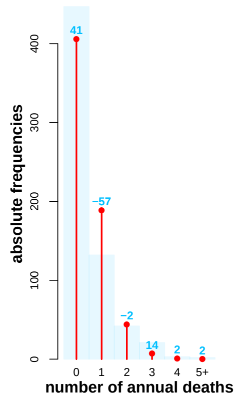


Figure: Comparison of observed and expected frequencies

Example (Negative binomial distribution; accidents in the factories)

Let X be the number of workers having an accident in munition factories in England during First World War (Greenwood and Yule 1920), n be the number of accidents, $m_{n,O}$ be the number of workers with particular number of accidents, $M = \sum_n m_{n,O} = 647$. **Question**: Calculate theoretical frequencies $m_{n,E}$.

Table: Observed and theoretical frequencies ($m_{n,O}$ and $m_{n,E}$) of workers with n accidents

n	0	1	2	3	4	≥ 5
$m_{n,O}$	447	132	42	21	3	2
$m_{n,E}$	446	134	44	15	5	3

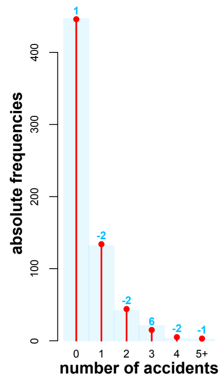


Figure: Comparison of observed and expected frequencies

Example (ZIP distribution; number of movements of a foetal lamb)

Let X be the number of movements of a foetal lamb in 240 five-second periods (Leroux and Puterman 1992), n be the number of movements, $m_{n,O}$ be the number of foetal lambs with particular number of movements. **Question**: Calculate theoretical frequencies $m_{n,E}$ using Poisson and ZIP distribution.

Table: Observed and theoretical frequencies ($m_{n,O}$ and $m_{n,E}$) of workers with n accidents

n	0	1	2	3	4	5	6	7
$m_{n,O}$	182	41	12	2	2	0	0	1
$m_{n,E}$ (Poisson)	168	60	11	1	0	0	0	0
$m_{n,E}$ (ZIP)	182	37	16	4	1	0	0	0

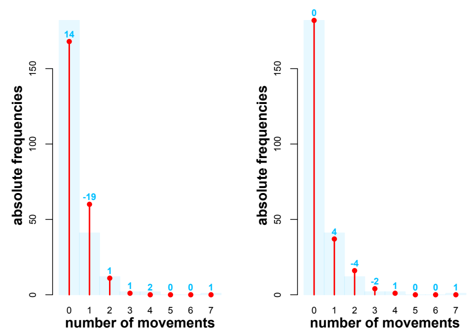


Figure: Comparison of observed and expected frequencies, Poisson (left), ZIP (right)

1. binomial distribution – example – **number of boys**:
   - Is the probability of number of boys in the families with 12 boys binomial?
   - Is the probability of having a boy in the family equal to 0.5?
2. multinomial distribution – example – **number of individuals with certain eye and hair colour**: Are the rows and columns of a contingency table independent?
   - Are the frequencies of individuals with certain eye colour (with levels blue, green, brown) independent of hair colour (with levels blond, light-brown, black, red)?

3. product-multinomial distribution: Are the vectors of frequencies the same in each row? Are the vectors of frequencies independent of the row index?
   - example – **number of individuals with certain socioeconomic status, political philosophy and affiliation** – Are the vectors of frequencies of individuals (D-Li, D-C, R-Li, R-C) the same for each level of socioeconomic status (high and low)?
   - example – **blood groups** – Is the distribution of the blood groups (0, A, B, AB) the same in Prague and Košice?
4. Poisson distribution:
   - example – **killing by horse kick** – Is the distribution of number of corps of soldiers with n annual deaths (killed by horse kicks) Poisson?
   - example – **accidents in the factories** – Is the distribution of number of workers having an accident Poisson?

Assignment **number of boys**:

1. Draw probability mass function of number of boys in the families with 12 children.
2. What are the probabilities of having n boys in the family ($n = 1, 2, \ldots, 12$)? What is the probability of having eight or more boys in the family? What is the probability of having five to seven boys in the family?

Assignment **killing by horse kick**:

1. Draw probability mass function of number of corps with n annual deaths (killed by horse kicks).
2. What are the probabilities of having n annual deaths ($n = 0, 1, 2, 3, 4, 5+$)? What is the probability of having one or less annual deaths?

Assignment **accidents in the factories**:

1. Draw probability mass function of number of workers having an accident.
2. What are the probabilities of having n accidents ($n = 0, 1, 2, 3, 4, 5+$)? What is the probability of having two or more accidents?

Assignment **number of boys**:

Calculate $\widehat{p}$ (the probability of having a boy in a family) and $\widehat{Var}[\widehat{p}]$ (the variance of probability of having a boy in a family).

Assignment **killing by horse kick**:

Calculate $\widehat{\lambda}$ (the mean number of annual deaths) and $\widehat{Var}[\widehat{\lambda}]$ (the variance of mean number of annual deaths).

Assignment **accidents in the factories**:

Calculate $\widehat{\lambda}$ (the mean number of accidents in the factories) and $\widehat{Var}[\widehat{\lambda}]$ (the variance of mean number of accidents in the factories).

Assignment **blood groups**:
In Prague and Košice, calculate $\widehat{\mathbf{p}}$ (the probabilities of having certain blood group in particular city) and $\widehat{Var[\widehat{\mathbf{p}}]}$ (the covariance matrix of probability of having certain blood group in particular city).

Assignment **eye and hair colour**:
Calculate $\widehat{\mathbf{p}}$ (the probabilities of having certain eye and hair colour) and $\widehat{Var[\widehat{\mathbf{p}}]}$ (the covariance matrix of probability of having certain eye and hair colour).

$1 \times J$ **contingency table of frequencies**

outcome 1	outcome 2	...	outcome J	sum
x_1	x_2	...	x_J	N

$1 \times J$ **contingency table of probabilities**

outcome 1	outcome 2	...	outcome J	sum
p_1	p_2	...	p_J	1

$2 \times J$ **contingency table of frequencies**

	outcome 1	outcome 2	...	outcome J	sum
row 1	x_{11}	x_{12}	...	x_{1J}	N_1
row 2	x_{21}	x_{22}	...	x_{2J}	N_2

$2 \times J$ **contingency table of probabilities**

	outcome 1	outcome 2	...	outcome J	sum
row 1	p_{11}	p_{12}	...	p_{1J}	$p_{1\bullet} \neq 1$
row 2	p_{21}	p_{22}	...	p_{2J}	$p_{2\bullet} \neq 1$

$K \times J$ **contingency table of frequencies**

	outcome 1	outcome 2	...	outcome J	sum
row 1	x_{11}	x_{12}	...	x_{1J}	N_1
row 2	x_{21}	x_{22}	...	x_{2J}	N_2
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
row K	x_{K1}	x_{K2}	...	x_{KJ}	N_K

$K \times J$ **contingency table of probabilities**

	outcome 1	outcome 2	...	outcome J	sum
row 1	p_{11}	p_{12}	...	p_{1J}	$p_{1\bullet} \neq 1$
row 2	p_{21}	p_{22}	...	p_{2J}	$p_{2\bullet} \neq 1$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
row K	p_{K1}	p_{K2}	...	p_{KJ}	$p_{K\bullet} \neq 1$

$1 \times J$ **contingency table of frequencies** ($\approx$ multinomial distribution)

outcome 1	outcome 2	...	outcome J	sum
x_1	x_2	...	x_J	N

$1 \times J$ **contingency table of probabilities** ($\approx$ multinomial distribution)

outcome 1	outcome 2	...	outcome J	sum
p_1	p_2	...	p_J	1

$2 \times J$ **contingency table of frequencies** ($\approx$ multinomial distribution)

	outcome 1	outcome 2	...	outcome J	sum
group 1	x_{11}	x_{12}	...	x_{1J}	N_1
group 2	x_{21}	x_{22}	...	x_{2J}	N_2

$2 \times J$ **contingency table of probabilities**

	outcome 1	outcome 2	...	outcome J	sum
group 1	$p_{1\|1}$	$p_{2\|1}$	...	$p_{J\|1}$	1
group 2	$p_{1\|2}$	$p_{2\|2}$	...	$p_{J\|2}$	1

$K \times J$ **contingency table of frequencies** ($\approx$ multinomial distribution)

	outcome 1	outcome 2	...	outcome J	sum
group 1	x_{11}	x_{12}	...	x_{1J}	N_1
group 2	x_{21}	x_{22}	...	x_{2J}	N_2
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
group K	x_{K1}	x_{K2}	...	x_{KJ}	N_K

$K \times J$ **contingency table of probabilities**

	outcome 1	outcome 2	...	outcome J	sum
group 1	$p_{1\|1}$	$p_{2\|1}$	...	$p_{J\|1}$	1
group 2	$p_{1\|2}$	$p_{2\|2}$	...	$p_{J\|2}$	1
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
group K	$p_{1\|K}$	$p_{2\|K}$	...	$p_{J\|K}$	1

	outcome 1	outcome 2	...	outcome J	sum
$\mathbf{x}_1$	1	0	...	0	1
$\mathbf{x}_2$	0	1		0	1
$\mathbf{x}_3$	0	1		0	1
$\mathbf{x}_4$	1	0	...	0	1
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
$\mathbf{x}_{N-1}$	0	0	...	1	1
$\mathbf{x}_N$	1	0	...	0	1
sum$=\mathbf{x}$	x_1	x_2	...	x_J	N

- sum of each row is one
- sum of all row sums is N
- sum of each column is x_j, where $j = 1, 2, \ldots, J$
- sum of all x_j, $j = 1, 2, \ldots, J$, is N
- $\mathbf{x} = \mathbf{n}$

	outcome 1	outcome 2	...	outcome J	sum
$\mathbf{x}_{k1}$	1	0	...	0	1
$\mathbf{x}_{k2}$	0	1		0	1
$\mathbf{x}_{k3}$	0	1		0	1
$\mathbf{x}_{k4}$	1	0	...	0	1
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
$\mathbf{x}_{k,N_k-1}$	0	0	...	1	1
$\mathbf{x}_{k,N_k}$	1	0	...	0	1
sum$=\mathbf{x}_k$	x_{k1}	x_{k2}	...	x_{kJ}	N_k

- sum of each row is one
- sum of all row sums is N_k
- sum of each column is x_{kj}, where $j = 1, 2, \ldots, J$
- sum of all x_{kj}, $j = 1, 2, \ldots, J$, is N_k
- $\mathbf{x}_k = \mathbf{n}_k$, where $k = 1, 2, \ldots, K$

Definition (normal distribution)

Random variable X is **normally distributed** with parameters μ and σ^2, i.e. $X \sim N(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ and density is defined as $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}, \sigma > 0$.

Definition (standardised normal distribution)

Random variable X is **normally distributed** with parameters $\mu = 0$ and $\sigma^2 = 1$, i.e. $X \sim N(0, 1)$, where $\boldsymbol{\theta} = (0, 1)^T$ and density is defined as $\phi(x) = f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \in \mathbb{R}$.

Parameter μ is called **mean** of X and σ^2 the **variance** of X.

Definition (bivariate normal distribution)

Random vector $(X, Y)^T$ is **normally distributed** with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, i.e. $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (\mu_1, \mu_2)^T \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

$\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$, $(x, y)^T \in \mathbb{R}^2$, $\mu_j \in \mathbb{R}^1$, $\sigma_j^2 > 0, j = 1, 2$, $\rho \in \langle -1, 1\rangle$; density is defined as

$$f(x, y) = \frac{1}{A} \exp\left\{-\frac{1}{B}\left\{\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right\}\right\},$$

where $A = 2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}$, $B = 2(1-\rho^2)$.

Definition (bivariate standardised normal distribution)

Random vector $(X, Y)^T$ is **normally distributed** with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, i.e. $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (0, 0)^T \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

$\boldsymbol{\theta} = (0, 0, 1, 1, \rho)^T$, $(x, y)^T \in \mathbb{R}^2$, $\rho \in \langle -1, 1\rangle$; density is defined as

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right\}.$$

Let $x = x_1$, $y = x_2$ and $\mathbf{x} = (x_1, x_2)^T$. Then the density of standardised bivariate normal distribution can be rewritten into matrix form:

$$f(\mathbf{x}) = \frac{1}{2\pi(\det(\boldsymbol{\Sigma}))^{1/2}} \exp\left\{-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right\}.$$

Let $(X_1, X_2, \ldots, X_k)^T \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{x}$ is k-dimensional vector, then the density is equal to

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}(\det(\boldsymbol{\Sigma}))^{1/2}} \exp\left\{-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right\}.$$

Marginal distributions of:

- bivariate normal distribution – $X_j \sim N(\mu_j, \sigma_j^2), j = 1, 2, \ldots, k$
- standardised bivariate normal distribution – $X_j \sim N(0, 1), j = 1, 2, \ldots, k$

Simulation of pseudo-random numbers from bivariate normal distribution:

1. let $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 1)$
2. then $(Y_1, Y_2)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $Y_1 = \sigma_1 X_1 + \mu_1$ and $Y_2 = \sigma_2(\rho X_1 + \sqrt{1-\rho^2}X_2) + \mu_2$

Example

Simulate pseudo-random numbers from bivariate normal distribution, where $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$.
(a) $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 1, \rho = 0$; (1) $n = 50$ and (2) $n = 1000$;
(b) $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 1, \rho = 0.5$; (1) $n = 50$ and (2) $n = 1000$;
(c) $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 1.2, \rho = 0.5$; (1) $n = 50$ and (2) $n = 1000$.

Probabilistic and Statistical Models
Bivariate normal models

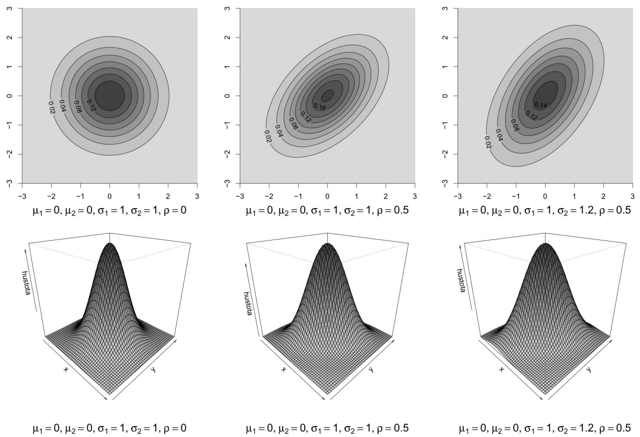


Figure: Joint density of three different bivariate normal distributions (column by column); contour plots superimposed by image plots (first row), 3D surface plot (second row); simulation study

Probabilistic and Statistical Models
Bivariate normal models

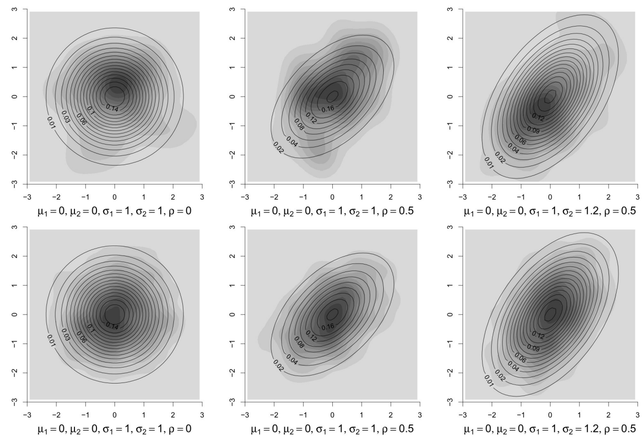


Figure: Joint density of three different bivariate normal distributions (column by column); $n = 50$ (first row), $n = 1000$ (second row); contour plots superimposed by image plots; simulation study

Probabilistic and Statistical Models
Mixture of two univariate and bivariate normal distribution

The mixture of two univariate normal distribution is defined as follows: $pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2)$, where $\boldsymbol{\theta} = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^T$.

The mixture of two bivariate normal distribution is defined as follows: $pN_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-p)N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\theta} = (p, \mu_{11}, \mu_{12}, \sigma_{11}^2, \sigma_{12}^2, \rho_1, \mu_{21}, \mu_{22}, \sigma_{21}^2, \sigma_{22}^2, \rho_2)^T$.

Probabilistic and Statistical Models
Different normal models – skewed, mesokurtic, platykurtic and leptokurtic

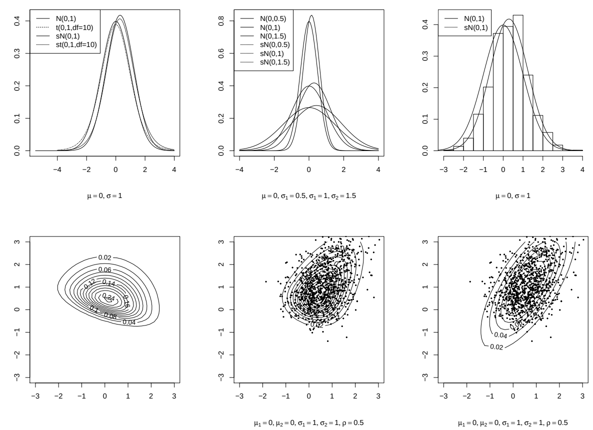


Figure: Densities of different normal and skewed normal distributions (first row, skewed normal indicated as ”sN”), densities of different bivariate skewed normal distributions (second row)

Probabilistic and Statistical Models
The mixture of two bivariate normal distribution

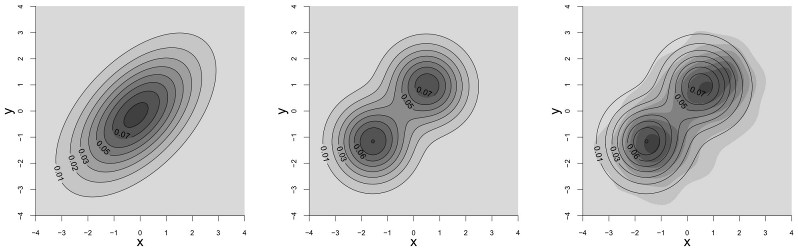


Figure: Joint density of bivariate normal distribution (left), density of the mixture of two bivariate normal distributions (middle), bivariate kernel density estimate superimposed by density of the mixture of two bivariate normal distributions (right); simulation study (contour plots superimposed by image plots)

Probabilistic and Statistical Models
Mixture of two univariate normal distribution

To express the **binormal distribution** formally, let B_i be (unobserved) iid *Bernoulli*(p) random variable, $p \in (0,1)$. If $B_i = 1$ then X_i is observed from $N(\mu_1, \sigma_1^2)$ distribution, otherwise it is observed from $N(\mu_2, \sigma_2^2)$. Thus, the distribution of X_i given by B_i is

$$X_i|(B_i = b_i) \sim \begin{cases} N(\mu_1, \sigma_1^2), & \text{if } b_i = 1, \\ N(\mu_2, \sigma_2^2), & \text{if } b_i = 0. \end{cases}$$

The **joint density** of (X_i, B_i) is therefore given by

$$f(x_i, b_i, \boldsymbol{\theta}) = f(x_i|b_i, \boldsymbol{\theta}) \Pr(B_i = b_i, p) \sim \begin{cases} \frac{p}{\sqrt{2\pi}\sigma_1} \exp(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}), & \text{if } b_i = 1, \\ \frac{1-p}{\sqrt{2\pi}\sigma_2} \exp(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}), & \text{if } b_i = 0, \end{cases}$$

where $\boldsymbol{\theta} = (p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)^T$, from which the marginal density of X_i is obtained as

$$f(x_i, \boldsymbol{\theta}) = \sum_{b_i \in \{0,1\}} f(x_i, b_i, \boldsymbol{\theta}) = f(x_i, 0, \boldsymbol{\theta}) + f(x_i, 1, \boldsymbol{\theta}).$$

Probabilistic and Statistical Models
Mixture of two univariate normal distribution

The binormal density function is a linear combination of the density functions given by $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributions.

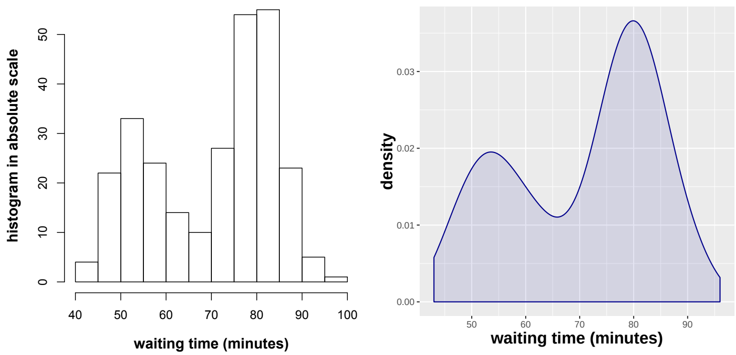


Figure: Mixture of two normal densities – data `faithful`

Probabilistic and Statistical Models
Binomial distribution

Jacob Bernoulli (1655–1705) – one of the founding fathers of probability theory.

Definition (binomial distribution)

Let N be number of independent identical (random) *Bernoulli trials* X_i, where $X_i = 1$ is a **success** (event occurred) and $X_i = 0$ is a **failure** (event did not occur), $i = 1, 2, \ldots, N$. Then **probability of success** $\Pr(X_i = 1) = p$ and **probability of failure** $\Pr(X_i = 0) = 1 - p$. Number of successes $X = \sum_{i=1}^N X_i$. The probability that random variable X is equal to $x = n$ (realisation) is defined as $\Pr(X = x) = \binom{N}{x} p^x (1-p)^{N-x}$, for $x = 0, 1, 2, \ldots, N$.
Expected value of X is defined as $E[X] = \sum_{x=0}^N x \Pr(X = x) = \sum_{x=0}^N x \binom{N}{x} p^x (1-p)^{N-x} = Np$.
Variance of X is defined as $Var[X] = \sum_{x=0}^N (x - E[X])^2 \Pr(X = x) = \sum_{x=0}^N (x - Np)^2 \binom{N}{x} p^x (1-p)^{N-x} = Np(1-p)$.

Reading: Random variable X is binomially distributed with parameters N and p, where $\theta = p$.
Notation: $X \sim Bin(N, p), \theta = p$
Do we need to change it? YES.
Why? Due to generalisation.

Equivalently, $\mathbf{X} \sim Bin\,(N, p, 1-p)$, where $\mathbf{X} = (X_1, X_2)^T$, $\boldsymbol{\theta} = (p, 1-p)^T$, X_1 is **number of successes**, $X_2 = N - X_1$ is **number of failures**, $X_1 \sim Bin\,(N, p)$ and $X_2 \sim Bin\,(N, 1-p)$.
Then

- $E[X_1] = Np$, $E[X_2] = N\,(1-p)$,
- $Var[X_2] = Np\,(1-p) = Var[X_1]$ is independent of p,
- $Cov\,[X_1, X_2] = -Np\,(1-p)$,
- $Cor\,[X_1, X_2] = -1$.

Finally, $\mathbf{n} = (n_1, n_2)^T$ and $\mathbf{p} = (p_1, p_2)^T$, $p_1 = p$ and $p_2 = 1 - p$.
Then $\boldsymbol{\theta} = \mathbf{p}$.

If each selection from a population of size N_{pop} is returned to the population, i.e. the sampling is with replacement, then, for each selection, the probability of selecting an individual with given characteristic is $p = M/N_{pop}$, where number of individuals with given characteristic is M (M means ”marked”) and the proportion can now be treated as a probability. Since the selections or ”trials” are mutually independent and number of trials N is fixed, *number of outcomes X having given characteristic in the sample* now has a **Binomial distribution**, denoted by $Bin(N, p)$.

Definition (binomial distribution)

If a random sample of size N is taken from the population of size N_{pop} with replacement and X is the number of individuals with a given characteristic in the sample, then X has a **binomial distribution** with probability mass function defined as $\Pr(X = x) = \binom{N}{x} p^x\,(1-p)^{N-x}$, where $x = 0, 1, 2, \ldots, N$.
Expected value of X is defined as $E\,[X] = Np$.
Variance of X is defined as $Var\,[X] = Np\,(1-p)$.

If we remove and individual chosen at random from the population of size N_{pop} and chose a second individual at random from the remainder, then the probability of getting an individual with given characteristic (M means ”marked”) is $(M-1)/(N_{pop}-1)$ if the first individual was with this characteristic and $M/(N_{pop}-1)$ if it was not. This is called sampling without replacement and the probability of choosing an individual with given characteristic changes with each selection. Then *number of outcomes X having given characteristic* now has a **Hypergeometric distribution**, denoted by $HypGeom(N, p)$.

Definition (hypergeometric distribution)

If a random sample of size N is taken from the population of size N_{pop} without replacement and X is the number of individuals with a given characteristic in the sample, then X has a **hypergeometric distribution** with probability mass function defined as $\Pr(X = x) = \binom{M}{x}\binom{N_{\text{pop}}-M}{N-x}/\binom{N_{\text{pop}}}{N}$, where $\max\{N + M - N_{\text{pop}}, 0\} \leq x \leq \min\{M, N\}$, but we usually have $x = 0, 1, 2, \ldots, N$.
Expected value of X is defined as $E[X] = Np$.
Variance of X is defined as $Var[X] = Np(1-p)r$, where $r = \frac{N_{\text{pop}}-N}{N_{\text{pop}}-1} = 1 - \frac{N-1}{N_{\text{pop}}-1} > 1 - f_s$, $f_s = N/N_{\text{pop}}$ is sampling fraction. (f_s can generally be neglected if $f_s < 0.1$ (or preferably $f_s < 0.05$) and we can then set $r = 1$)

We see then that if f_s can be ignored, we can **approximate sampling without replacement by sampling with replacement**, and **approximate the hypergeometric distribution by the binomial distribution**.

Definition (multinomial distribution)

Let N be number of independent identical (random) trials and in each of them $J \geq 2$ distinct possible outcomes can occur, where $X_{ji} = 1$ is a **success** (event occurred) and $X_{ji} = 0$ is a **failure** (event did not occur), $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, J$.
Number of successes $X_j = \sum_{i=1}^{N} X_{ji}$, $N = \sum_{j=1}^{J} X_j$. Then **probability of success** of j-th outcome in i-th trial is equal to $\Pr(X_{ji} = 1) = p_j$ (**cell probabilities**) and **probability of failure** in j-th trial is equal to $\Pr(X_{ji} = 0) = 1 - p_j$. Let $\mathbf{X} = (X_1, X_2, \ldots, X_J)^T$. The probability that random variables X_j are equal to $x_j = n_j$ is defined as

$$\Pr(X_1 = x_1, \ldots, X_J = x_J) = \frac{N!}{\prod_j x_j!}\prod_{j=1}^{J} p_j^{x_j}.$$

Expected value of $\mathbf{X}$ is a vector defined as $E[\mathbf{X}] = N\mathbf{p}$.
Covariance matrix of $\mathbf{X}$ is defined as

$$Var[\mathbf{X}] = N\left(\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T\right),$$

where

$$(Var[\mathbf{X}])_{ij} = \begin{cases} Np_j(1-p_j) & \text{if } i = j \\ -Np_ip_j & \text{if } i \neq j \end{cases}$$

Marginal distributions are binomial, i.e. $X_j \sim Bin(N, p_j)$.
Then

- $E[X_j] = Np_j$,
- $Var[X_j] = Np_j(1-p_j)$
- $Cov[X_i, X_j] = -Np_ip_j$
- $Cor[X_i, X_j] = (-p_ip_j)/\sqrt{p_i(1-p_i)p_j(1-p_j)}$

Reading: Random vector $\mathbf{X}$ is multinomially distributed with parameters N and $\mathbf{p}$, where $\theta = \mathbf{p}$.
Notation: $\mathbf{X} \sim Mult_J(N, \mathbf{p})$.
If $J = 2$, then $Bin(N, p) \approx Mult_2(N, \mathbf{p})$
Realisation of one trial $\mathbf{x}_{ji}$ could be $(1, 0, \ldots, 0)^T$ or $(0, 1, \ldots, 0)^T$.

Example (number of individuals with certain blood type)

Number of individuals $\mathbf{X} = (X_1, X_2, X_3, X_4)^T$ with certain blood group is multinomially distributed following Hardy-Wienberg equilibrium, i.e. $\mathbf{X} = (X_1, X_2, X_3, X_4)^T \sim Mult_4(N, \mathbf{p})$, where $N = 500$ (Katina et al. 2015). Calculate theoretical frequencies $n_{j,E}$.

attributes (groups)	0	A	B	AB
$n_{j,O}$	209	184	81	26
$n_{j,E}$	210	183	80	27

Example (number of individuals with certain socioeconomic status, political philosophy and political affiliation)

Number of individuals $X_1, \ldots, X_8$ with socioeconomic status, political philosophy and political affiliation is multinomially distributed, i.e. $\mathbf{X} = (X_1, \ldots, X_8)^T \sim Mult_8(N, \mathbf{p})$, where $\mathbf{p} = (p_1, p_2, \ldots, p_8)^T$ and $N = 500$ (Christensen 1990, modified). Calculate (a) $Var[X_1]$, (b) $Var[X_4]$, (c) $Cov[X_1, X_4]$ and (d) $Cor[X_1, X_4]$.

Table: 2×4 contingency table of probabilities p_j

	D-C	D-Li	R-C	R-Li	total
H	0.12	0.12	0.12	0.04	0.4
Lo	0.18	0.18	0.18	0.06	0.6
total	0.30	0.30	0.30	0.10	1.0

Notation: (1) socioeconomic status (high – H, low – Lo), (2) political philosophy (democrat – D, republican – R) a (3) political affiliation (conservative – C, liberal – Li). Then X_1 (H-D-C), X_2 (H-D-Li), X_3 (H-R-C), X_4 (H-R-Li), X_5 (Lo-D-C), X_6 (Lo-D-Li), X_7 (Lo-R-C) and X_8 (Lo-R-Li).
Solution:
$Var[X_1] = 500 \times 0.12 \times (1 - 0.12) = 52.8$
$Var[X_4] = 500 \times 0.04 \times (1 - 0.04) = 19.2$
$Cov[X_1, X_4] = -500 \times 0.12 \times 0.04 = -2.4$
$Cor[X_1, X_4] = -2.4/\sqrt{52.8 \times 19.2} = -0.075$

What are the expected frequencies?

Table: 2×4 contingency table of frequencies X_j

	D-C	D-Li	R-C	R-Li
H	60	60	60	20
Lo	90	90	90	30

Definition (multi-hypergeometric distribution)

Suppose we have k subpopulations of sizes M_j, where $j = 1, 2, \ldots, k$, and $\sum_{j=1}^k M_j = N_{\text{pop}}$, the total population size. Let $p_j = M_j / N_{\text{pop}}$. A simple random sample of size N is taken from the population yielding X_j from the j-th subpopulation. $\mathbf{X} = (X_1, X_2, \ldots, X_k)^T$ has **multi-hypergeometric distribution**. Then joint probability mass function of $\mathbf{X}$ is defined as $f(\mathbf{x}) = \Pr(\mathbf{X} = \mathbf{x}) = \prod_{j=1}^k \binom{M_j}{x_j} / \binom{N_{\text{pop}}}{N}$, where $0 \le x_j \le \min\{M_j, N\}$ and $N = \sum_{j=1}^k x_j$.

Since we can add the subpopulations together we see that the marginal distribution of an X_j is also hypergeometric, with two subpopulations M_j and $N - M_j$, namely $f_j(x_j) = \binom{M_j}{x_j}\binom{N_{\text{pop}}-M_j}{N-x_j}/\binom{N_{\text{pop}}}{N}$.

In a similar fashion we see that the probability function of $X_1 + X_2$ is the multi-hypergeometric distribution, namely $f_{12}(x_1, x_2) = \binom{M_1+M_2}{x_1+x_2}\binom{N_{\text{pop}}-M_1-M_2}{N-x_1-x_2}/\binom{N_{\text{pop}}}{N}$.

Additionally, $Var[X_j] = Np_j(1-p_j)r$, where $r = (N_{\text{pop}} - N)/(N_{\text{pop}} - 1)$, and $Var[X_1 + X_2] = Nr(p_1 + p_2)(1 - p_1 - p_2)$. Finally, the covariance of X_1 and X_2 is equal to $Cov[X_1, X_2] = \frac{1}{2}(Var[X_1 + X_2] - Var[X_1] - Var[X_2]) = -rNp_1p_2$.

We then find that if $q_j = 1 - p_j$, then $Var[X_1 - X_2] = Var[X_1] + Var[X_2] - 2Cov[X_1, X_2] = rN[p_1q_1 + p_2q_2 - 2p_1p_2] = rN\left[p_1 + p_2 - (p_1 - p_2)^2\right]$.

Suppose we have a population of N_{pop} people and a sample of size N is chosen at random without replacement. Each selected person is asked two questions to each of which they answer *yes* (1) or *no* (2), so that p_{12} is the proportion answering *yes* to the first question and *no* to the second, p_{11} is the proportion answering *yes* to both questions, and so forth. Then the proportion answering *yes* to the first question is $p_1 = p_{11} + p_{12}$ and the proportion answering *yes* to the second question is $p_2 = p_{11} + p_{21}$. Let X_{ij} $(i, j = 1, 2)$ be the number observed in the sample in the category with probability p_{ij}, let $X_1 = X_{11} + X_{12}$ the number answering *yes* to the first question, and let $X_2 = X_{11} + X_{21}$ be the number answering yes to the second question. The interest is to compare p_1 and p_2 but p_{12} is often ignored (and p_{21} as well).

The four variables X_{ij} have a multi-hypergeometric distribution, and $\frac{X_1}{N} - \frac{X_2}{N} = \frac{X_1 - X_2}{N} = \frac{X_{12} - X_{21}}{N} = \frac{X_{12}}{N} - \frac{X_{21}}{N}$

$$E\left[\frac{X_1}{N} - \frac{X_2}{N}\right] = p_1 - p_2 = p_{12} - p_{21},$$

Finally, $Var\left[\frac{X_1}{N} - \frac{X_2}{N}\right] = \frac{1}{N^2}Var[X_1 - X_2] = \frac{1}{N^2}(Var[X_1] + Var[X_2] - 2Cov[X_1, X_2]) = r\frac{1}{N}[p_1q_1 + p_2q_2 - 2p_1p_2] = r\frac{1}{N}\left[p_1 + p_2 - (p_1 - p_2)^2\right]$.

If we can approximate sampling without replacement by sampling with replacement, we can set $r = 1$ above, and the **multi-hypergeometric distribution can be replaced by the multinomial distribution**.

The Multinomial distribution also arises when we have N fixed Bernoulli trials but with k possible outcomes rather than just two, as with the binomial distribution.

Definition (product-multinomial distribution)

Let N_k be number of independent identical (random) trials and in each of them $J \geq 2$ distinct possible outcomes can occur, where $X_{kji} = 1$ is a **success** (event occurred) and $X_{kji} = 0$ is a **failure** (event did not occur), $i = 1, 2, \ldots, N_k$, $k = 1, 2, \ldots, K$, $j = 1, 2, \ldots, J$. Number of successes $X_{kj} = \sum_{i=1}^{N_k} X_{kji}$ and $\sum_{k=1}^{K} N_k = N$. Then **probability of success** of kj-th outcome in i-th trial is equal to $\Pr(X_{kji} = 1) = p_{kj}$ (**cell probabilities**) and **probability of failure** of kj-th outcome in i-th trial is equal to $\Pr(X_{kji} = 0) = 1 - p_{kj}$. Let $\mathbf{X}_k = (X_{k1}, X_{k2}, \ldots, X_{kJ})^T$ be multinomially distributed with parameters N_k and $\mathbf{p}_k$, i.e. $\mathbf{X}_k \sim Mult_J(N_k, \mathbf{p}_k)$, where $\boldsymbol{\theta}_k = \mathbf{p}_k$ a $\mathbf{p}_k = (p_{k1}, p_{k2}, \ldots, p_{kJ})^T$. Let realisations of $\mathbf{X}_k$ be $\mathbf{x}_k$. Then $x_{kj} = n_{kj}$ and $\mathbf{n}_k = (n_{k1}, n_{k2}, \ldots, n_{kJ})^T$. Additionally, $\mathbf{X}_k$ are independent.

The probability that random variables X_{kj} are equal to $x_{kj} = n_{kj}$ (for all j and k) is defined as

$$\Pr(X_{kj} = x_{kj}, \forall k, j) = \prod_{k=1}^{K} \Pr(X_{kj} = x_{kj}, \forall j).$$

The probability that random variables X_{kj} are equal to $x_{kj} = n_{kj}$ (for all j) is defined as

$$\Pr(X_{kj} = x_{kj}, \forall j) = \left(N_k! / \prod_{j=1}^{J} x_{kj}!\right) \prod_{j=1}^{J} p_{kj}^{x_{kj}}.$$

Then

$$\Pr(X_{kj} = x_{kj}, \forall k, j) = \prod_{k=1}^{K} \left(\left(N_k! / \prod_{j=1}^{J} x_{kj}!\right) \prod_{j=1}^{J} p_{kj}^{x_{kj}} \right).$$

Reading: Random matrix $\mathbf{X}$ is product-multinomially distributed with parameters $\mathbf{N} = (N_1, N_2, \ldots, N_K)^T$ and $\mathbf{p}$ with the rows $\mathbf{p}_k$, where $\boldsymbol{\theta}_k = \mathbf{p}_k, k = 1, 2, \ldots, K$.
Notation: $\mathbf{X} \sim ProdMult_K(\mathbf{N}, \mathbf{p})$.
If $K = 1$, then $Mult_J(N, \mathbf{p}) \approx ProdMult_1(N, \mathbf{p})$
Realisation of one trial $\mathbf{x}_{kij}$ could be $(1, 0, \ldots, 0)^T$ or $(0, 1, \ldots, 0)^T$.
Then

- **expected frequencies** are equal to $N_k p_{kj}$,
- within each $\mathbf{X}_k$, **variances** $Var[X_{kj}]$, **covariances** $Cov[X_{kj}, X_{ki}]$ and **correlations** $Cor[X_{kj}, X_{ki}]$ are calculated as for multinomial distribution,
- between $\mathbf{X}_k$, e.g. $Cov[\mathbf{X}_1, \mathbf{X}_2]$, $k = 1, 2$, are zeroes due to independence of $\mathbf{X}_k$

Example (number of individuals with certain socioeconomic status, political philosophy and political affiliation)

Number of individuals $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^T$ with socioeconomic status, political philosophy and political affiliation is product-multinomially distributed, i.e. $\mathbf{X} \sim ProdMult_2(\mathbf{N}, \mathbf{p})$, where $\mathbf{X}_1 = (X_{11}, X_{12}, X_{13}, X_{14})^T$ are number of individuals with high socioeconomic status, $\mathbf{X}_2 = (X_{21}, X_{22}, X_{23}, X_{24})^T$ number of individuals with low socioeconomic status, $\mathbf{p}_k = (p_{1|k}, p_{2|k}, \ldots, p_{J|k})^T$, $p_{kj} = p_{j|k} = \frac{n_{jk}}{n_k}$, $k = 1, 2$, $\mathbf{N} = (N_1, N_2)^T$, $N_1 = 200$, $N_2 = 300$ (Christensen 1990, modified). Calculate (a) probabilities $p_{j|k}$, (b) expected frequencies, (c) $Var[X_{4|1}]$, (d) $Cov\left[X_{1|2}, X_{4|2}\right]$ and (e) $Cov\left[X_{1|1}, X_{4|2}\right]$.

Notation: (1) socioeconomic status (high – H, low – Lo), (2) political philosophy (democrat – D, republican – R) a (3) political affiliation (conservative – C, liberal – Li). Then X_1 (H-D-C), X_2 (H-D-Li), X_3 (H-R-C), X_4 (H-R-Li), X_5 (Lo-D-C), X_6 (Lo-D-Li), X_7 (Lo-R-C) and X_8 (Lo-R-Li).

Solution:

Table: 2×4 contingency table of probabilities $p_{j|k}$

	D-C	D-Li	R-C	R-Li	total
H	0.3	0.3	0.3	0.1	1.0
Lo	0.3	0.3	0.3	0.1	1.0

Table: 2×4 contingency table of frequencies n_{kj}

	D-C	D-Li	R-C	R-Li	total
H	60	60	60	20	200
Lo	90	90	90	30	300

$Var[X_{4|1}] = 200 \times 0.1 \times (1 - 0.1) = 18.$
$Cov\left[X_{1|2}, X_{4|2}\right] = -300 \times 0.3 \times 0.1 = -9,$
$Cov\left[X_{1|1}, X_{4|2}\right] = 0$, due to the independence of $\mathbf{X}_1$ and $\mathbf{X}_2$.

Definition (Poisson distribution)

Let X be random variable characterised by Poisson distribution, i.e. $X \sim Poiss(\lambda)$, where $\theta = \lambda$. Then

$$\Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, \ldots,$$

where $x = n$ is realisation of X. Then $E[X] = \lambda$ and $Var[X] = \lambda$.

Binomial distribution can be approximated by Poisson distribution if $N \to \infty$, $p \to 0$ and $\lambda_N = Np \to \lambda$, where $X \sim Poiss(\lambda)$.

Poisson distribution can be approximated by χ^2 distribution if $N \to \infty$, $p \to 0$ and $\lambda_N = Np \to \lambda$ and $\Pr(X \leq y) = \Pr(\chi^2_{2(1+y)} \leq 2\lambda)$, where $X \sim Poiss(\lambda)$.

Example (Poisson distribution; number of car accidents per week)

Having 50 million people driving car independently in Italy next week, the probability of **car crash deaths** (road traffic deaths) is 0.000002 (**death rate**), where number of deaths X is distributed binomially, i.e. $Bin(50\,\text{mil}, 0.000002)$ or $Poiss(50\,\text{mil} \times 0.000002) \approx Poiss(100)$.

Example (Poisson distribution, three types of accidents)

Let n_1 be number of **car crash deaths**, n_2 be number of **airplane crash deaths**, n_3 be number of **train crash deaths** in Italy next week. Then Poisson model with parameters λ_1, λ_2 a λ_3 for independent Poisson random variables X_1, X_2 a X_3 is defined as $X_1 + X_2 + X_3 \sim Poiss(\lambda_1 + \lambda_2 + \lambda_3)$.

Generalising this example we get

$$X_1 + X_2 + \ldots + X_J \sim Poiss\left(\lambda_1 + \lambda_2 + \ldots + \lambda_J\right).$$

Multinomial distribution can be approximated by Poisson distribution

$$(X_1 + X_2 + \ldots + X_J)\,|N \sim Mult_J\,(N, p_1, p_2, \ldots, p_J)\,,$$

where $N = \sum_j X_j$ and $p_j = \frac{\lambda_j}{\sum_j \lambda_j}, j = 1, 2, \ldots, J$. If $X_j, j = 1, 2, \ldots, J$ are independent, $X_j \sim Poiss\,(\lambda_j)$, where $E[X_j] = \lambda_j$, then conditional probability, that all $X_j = x_j$ fixing (conditioning on) $N = \sum_j X_j$ is equal to

$$\Pr\left[\mathbf{X} = \mathbf{x} | \sum_j X_j = N\right] = \frac{\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_J = x_J)}{\Pr(\sum_j X_j = N)}$$

$$= \frac{\prod_j \frac{\lambda_j^{x_j} e^{-\lambda_j}}{x_j!}}{\frac{\lambda^N e^{-\lambda}}{N!}} = \frac{N! e^{-\lambda} \prod_j \lambda_j^{x_j}}{e^{-\lambda} \prod_j \lambda^x \prod_j x_j!}$$

$$= \frac{N!}{\prod_j x_j!} \prod_j \left(\frac{\lambda_j}{\lambda}\right)^{x_j}, \text{ where } p_j = \frac{\lambda_j}{\lambda}.$$

Definition (cumulative distribution function)

Let X be random variable. The **cumulative distribution function** of X is defined as

$$F_X(x) = \Pr(X \leq x).$$

for all $x \in \mathbb{R}$, where $\mathbb{R}$ is called a *domain* and with $\langle 0, 1\rangle$ as *counterdomain*.

Properties of cumulative distribution function:

1. $F_X(-\infty) = \lim_{x\to-\infty} F_X(x) = 0$, and $F_X(\infty) = \lim_{x\to\infty} F_X(x) = 1$.
2. $F_X(x)$ is a monotone, nondecreasing function, i.e. $F_X(a) \leq F_X(b)$ for $a < b$.
3. $F_X(x)$ is *right continuous* in each argument, i.e. $\lim_{0<h\to 0} F(x+h) = F(x)$.

Definition (joint cumulative distribution function)

Let $X_1, X_2, \ldots, X_k$ be k random variables. The **joint cumulative distribution function** of $X_1, X_2, \ldots, X_k$ is defined as

$$F_{X_1,X_2,\ldots,X_k}(x_1, x_2, \ldots, x_k) = \Pr(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k)$$

for all $(x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$, where $\mathbb{R}^k$ is called a *domain* and with $\langle 0, 1\rangle$ as *counterdomain*.

Properties of bivariate cumulative distribution function:

1. $F_{XY}(-\infty, y) = \lim_{x\to-\infty} F_{XY}(x, y) = 0$ for $\forall y$, $F_{XY}(x, -\infty) = \lim_{y\to-\infty} F_{XY}(x, y) = 0$ for $\forall x$, and $\lim_{x,y\to\infty} F_{XY}(x, y) = F_{XY}(\infty, \infty) = 1$.
2. If $x_1 < x_2$ and $y_1 < y_2$, then $\Pr(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F_{XY}(x_2, y_2) - F_{XY}(x_2, y_1) - F_{XY}(x_1, y_2) + F_{XY}(x_1, y_1) \geq 0$
3. $F_{XY}(x, y)$ is *right continuous* in each argument, i.e. $\lim_{0<h\to 0} F_{XY}(x+h, y) = \lim_{0<h\to 0} F_{XY}(x, y+h) = F_{XY}(x, y)$.

Definition (marginal cumulative distribution functions)

If $F_{X_1,X_2,\ldots,X_k}(x_1, x_2, \ldots, x_k)$ is joint cumulative distribution function of $X_1, X_2, \ldots, X_k$, then the cumulative distribution functions $F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_k}(x_k)$ are called **marginal cumulative distribution functions**.

Definition (marginal cumulative distribution functions)

If $F_{X,Y}(x, y)$ is joint cumulative distribution function of X, Y, then the cumulative distribution functions $F_X(x)$ and $F_Y(y)$ are called **marginal cumulative distribution functions**.

Remark: $F_X(x) = F_{XY}(x, \infty)$ and $F_Y(y) = F_{XY}(\infty, y)$, i.e. knowledge of joint cumulative distribution function of X and Y implies knowledge of the two marginal cumulative distribution functions.

Definition (joint discrete random variable)

The k-dimensional random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)^T$ is defined to be a k**-dimensional discrete random vector** if it can assume values only at a countable number of points $(x_1, x_2, \ldots, x_k)^T \in \mathbb{R}^k$. We also say that the random variables $X_1, X_2, \ldots, X_k$ are **joint(ly) discrete random variables**.

Definition (joint discrete density function ≈ probability mass function)

If $\mathbf{X} = (X_1, X_2, \ldots, X_k)^T$ is k-dimensional discrete random vector, then the **joint discrete density function** of $\mathbf{X}$ is defined as

$$f_{X_1,X_2,\ldots,X_k}(x_1, x_2, \ldots, x_k) = \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$$

for all $(x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$, and is defined to be 0 otherwise.

Remark: $\sum f_{X_1,X_2,\ldots,X_k}(x_1, x_2, \ldots, x_k) = 1$, where the summation is over all possible values of $X_1, X_2, \ldots, X_k$.

Definition (marginal discrete density functions ≈ probability mass function)

If X and Y are jointly discrete random variables, then $f_X(x)$ and $f_Y(y)$ are called **marginal discrete density functions**. More generally, $X_{j_1}, \ldots, X_{j_m}$ be any subset of jointly discrete random variables $X_1, X_2, \ldots, X_k$, then $f_{X_{j_1},\ldots,X_{j_m}}(x_{j_1}, \ldots, x_{j_m})$ is also called a marginal density of m-dimensional random vector $(X_{j_1}, \ldots, X_{j_m})^T$.

Remark: If $X_1, X_2, \ldots, X_k$ are jointly discrete random variables, then any marginal discrete density can be found from joint density, but not conversely. E.g. if X and Y are jointly discrete random variables with values $(x_i, y_j), i = 1, 2, \ldots, k, j = 1, 2, \ldots, k$, then

$$f_X(x_i) = \sum_j f_{XY}(x_i, y_j),$$

where the summation is over all y_j for the fixed x_i.

Definition (joint continuous random variable and density)

The k-dimensional random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)^T$ is defined to be a k**-dimensional continuous random vector** if and only if there exists a function $f_{X_1,X_2,\ldots,X_k}(x_1, x_2, \ldots, x_k) \geq 0$ such that

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} f_{X_1,\ldots,X_k}(u_1, \ldots, u_k) du_1, \ldots, du_k$$

for all $(x_1, x_2, \ldots, x_k)^T \in \mathbb{R}^k$. Function $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k)$ is defined to be **joint continuous density function**.

As in one dimensional case, joint continuous (probability) density function has two properties:

1. $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) \geq 0$, and
2. $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) dx_1, \ldots, dx_k = 1$.

Definition (marginal continuous density functions)

If X and Y are jointly continuous random variables, then $f_X(x)$ and $f_Y(y)$ are called **marginal continuous density functions**. More generally, $X_{j_1}, \ldots, X_{j_m}$ be any subset of jointly continuous random variables $X_1, X_2, \ldots, X_k$, then $f_{X_{j_1},\ldots,X_{j_m}}(x_{j_1}, \ldots, x_{j_m})$ is also called a marginal density of m-dimensional random vector $(X_{j_1}, \ldots, X_{j_m})^T$.

Remark: If $X_1, X_2, \ldots, X_k$ are jointly continuous random variables, then any marginal continuous density can be found from joint density, but not conversely. E.g. if X and Y are jointly continuous random variables, then

$$f_X(x) = \frac{\partial F_X(x)}{\partial x} = \frac{\partial}{\partial x}\left[\int_{-\infty}^{x}\left(\int_{-\infty}^{\infty} f_{XY}(u, y)dy\right)du\right] = \int_{-\infty}^{\infty} f_{XY}(x, y)dy$$

and

$$f_Y(y) = \frac{\partial F_Y(y)}{\partial y} = \frac{\partial}{\partial y}\left[\int_{-\infty}^{y}\left(\int_{-\infty}^{\infty} f_{XY}(x, u)dx\right)du\right] = \int_{-\infty}^{\infty} f_{XY}(x, y)dx.$$

Definition (conditional discrete density function)

Let X and Y be jointly discrete random variables with joint discrete density function $f_{XY}(x,y)$. The **conditional discrete density function** of Y given $X=x$ is defined as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)},$$

if $f_X(x) > 0$.

Definition (conditional discrete cumulative distribution function)

Let X and Y be jointly discrete random variables, the **discrete cumulative distribution function** of Y given $X=x$ is defined to be $F_{Y|X}(y|x) = \Pr[Y \leq y, X = x]$ for all $f_X(x) > 0$.

Remark: $F_{Y|X}(y|x) = \sum_{j:y_j \leq y} f_{Y|X}(y|x)$.

Definition (conditional continuous density function)

Let X and Y be jointly continuous random variables with joint continuous density function $f_{XY}(x,y)$. The **conditional continuous density function** of Y given $X=x$ is defined as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)},$$

if $f_X(x) > 0$.

Definition (conditional continuous cumulative distribution function)

Let X and Y be jointly continuous random variables, the **conditional continuous cumulative distribution function** of Y given $X=x$ is defined as $F_{Y|X}(y|x) = \Pr[Y \leq y, X = x]$ for all $f_X(x) > 0$.

Remark: $F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(u|x)du$.

We can also write the following:
$\int_{-\infty}^{\infty} f_{Y|X}(y|x)dy = \int_{-\infty}^{\infty} \frac{f_{XY}(x,y)}{f_X(x)}dy = \frac{1}{f_X(x)} \int_{-\infty}^{\infty} f_{XY}(x,y)dy = \frac{f_X(x)}{f_X(x)} = 1.$

Example (joint normal density)

Prove that the function

$$f_{XY}(x,y) = \frac{1}{A}\exp\left\{-\frac{1}{B}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right\}$$

where $A = 2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}$, $B = 2\left(1-\rho^2\right)$, has the following property $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1$. To simplify the integral, you shall substitute $u = (x-\mu_X)/\sigma_X$ and $v = (y-\mu_Y)/\sigma_Y$, and then $w = \frac{u-\rho v}{\sqrt{1-\rho^2}}$ and $dw = \frac{du}{\sqrt{1-\rho^2}}$.

Theorem (marginal normal density)

If (X, Y) has a bivariate normal distribution, then the marginal distributions of X and Y are univariate normal distributions, i.e. X is normally distributed with mean μ_X and variance σ_X^2, and Y is normally distributed with mean μ_Y and variance σ_Y^2.

Example (marginal normal density)

Prove above mentioned theorem, e.g. for $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$ and substituting $v = (y-\mu_Y)/\sigma_Y$.

Probabilistic and Statistical Models
Conditional normal density

Theorem (conditional normal density)

If random vector $(X, Y)^T$ has a bivariate normal distribution, then the conditional distributions of Y given $X = x$ is normal with mean $\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ and variance $\sigma_Y^2(1-\rho^2)$ and density

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}}$$

$$\exp\left\{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left(y - \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right)\right\}.$$

Example (conditional normal density)

Prove above mentioned theorem using joint and marginal normal densities, i.e. prove that $f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$.

Probabilistic and Statistical Models
Stochastic independence

Definition (stochastic independence)

Let $(X_1, X_2, \ldots, X_k)^T$ be a k-dimensional random vector. $X_1, X_2, \ldots, X_k$ are defined to be **stochastically independent** if and only if

$$F_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \prod_{j=1}^{k} F_{X_j}(x_j) \text{ for all } x_1,\ldots,x_k.$$

Definition (stochastic independence)

Let $(X_1, X_2, \ldots, X_k)^T$ be a k-dimensional random vector. $X_1, X_2, \ldots, X_k$ are defined to be **stochastically independent** if and only if

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \prod_{j=1}^{k} f_{X_j}(x_j) \text{ for all } x_1,\ldots,x_k.$$

Remark: Often the word ”stochastically” is omitted.

Probabilistic and Statistical Models
Assignments in R

Assignment **number of individuals with certain socioeconomic status, political philosophy and affiliation**:

1. What is the number of all 2×4 contingency table with $N = 50$?
   $\binom{n+k-1}{k} = \binom{57}{8} = \binom{57}{49} = 1652411475$

```
choose(57,49)
choose(57,8)
```

2. What is the probability of getting the following 2×4 contingency table?

	D-C	D-Li	R-C	R-Li
H	5	7	6	4
Lo	8	7	10	3

$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_8 = x_8) =$
$\frac{50!}{5!7!4!6!8!7!3!10!}0.12^5 0.12^7 0.04^4 0.12^6 0.18^8 0.18^7 0.06^3 0.18^{10} =$
2.332506×10^{-6}

```
n <- c(5,7,6,4,8,7,10,3)
p <- c(.12,.12,.12,.04,.18,.18,.18,.06)
dmultinom(x=n, prob=p)  #  2.332506e-06
```


Probabilistic and Statistical Models
Assignments in R

Assignment **number of individuals with certain socioeconomic status, political philosophy and affiliation**:

1. What is the most probable 2×4 contingency table and what is the probability of getting it?

	D-C	D-Li	R-C	R-Li
H	6	6	6	2
Lo	9	9	9	3

$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_8 = x_8) =$
$\frac{50!}{6!6!2!6!9!9!3!9!}0.12^6 0.12^6 0.12^6 0.04^2 0.18^9 0.18^9 0.18^9 0.06^3 =$
1.020471×10^{-5}
$4.375\times$ more than in (2)

```
n <- c(6,6,6,2,9,9,9,3)
p <- c(.12,.12,.12,.04,.18,.18,.18,.06)
dmultinom(x=n, prob=p) #  1.020471e-05
```

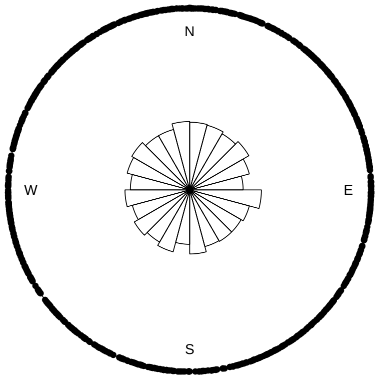
2. Draw probability mass function of number of possible 2×4 contingency tables with $N = 50$.

Example (histogram on a circle, rose diagram)

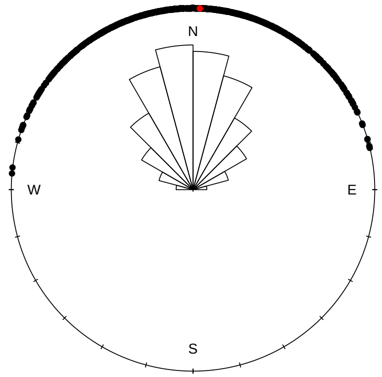
A **wind rose** is a graphic tool used by meteorologists to give a succinct view of how wind speed and direction are typically distributed at a particular location. In statistics, it is called **bivariate histogram**. Visualise in R wind rose of wind speed X_s in m/s (for a reference $1\,\text{m/s} = 3.6\,\text{km/h}$) and wind direction X_d in dgr of simulated data:

(A) $X_d \sim Unif(a, b)$, where $a = 0$ and $b = 360$, $X_s \sim Gamma(\lambda, k)$, where $\lambda = 50$ and $k = 1$ ($Gamma(\lambda, 1) \approx Exp(\lambda)$), $n = 1000$.
(B) $X_d \sim WN(\mu, \rho)$, where $\mu = 0$ and $\rho = \exp(-\sigma^2/2)$, $\sigma = 0.5$, $X_s \sim Gamma(\lambda, k)$, where $\lambda = 50$ and $k = 1$ ($Gamma(\lambda, 1)$), $n = 1000$.

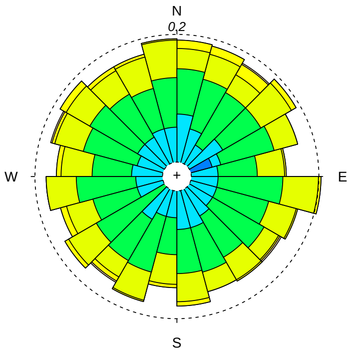
Use `library(circular)` and function `windrose()`. To visualise wind speed use also function `topo.colors(k)`. Be careful with colour scaling of k ordered intervals of wind speed. Visualise also rose diagrams, data and averages of wind direction (the latter when appropriate) and compare it with wind rose (orientation, scaling, etc.).



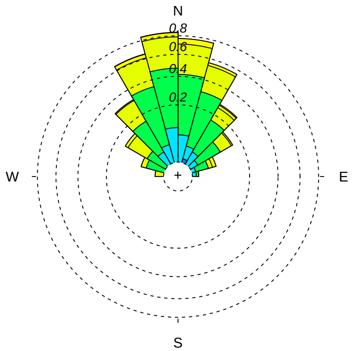
uniform distribution



wrapped normal distribution



uniform distribution of wind direction



wrapped normal distribution of wind direction

A common phenomenon is **the *arrival* or *occurrence* of an event at a time t independently of the time of previous occurrence of the events** – **events on nonoverlapping time intervals are mutually independent**. In addition, the average rate of arrivals is constant. The **Poisson probability mass function** (pmf) is a good model for **the number of arrivals in an interval** t and in general we call it a **Poisson process**. Typical applications include *occurrence of earthquakes*. As we increase the rate, the pmf would be more and more like a **normal distribution**.

We are interested in determining:

- the **pmf of the number arrivals in a time interval** t,
- **the probability density function (pdf) of the arrival time of the kth occurrence** (e.g. $k = 0, k = 1, k > 1$), and
- **the pdf of the time interval between arrivals of successive occurrences** (interarrival time).

Note: This process refers to arrivals on a continuous line. For many applications, this line is **time**, but for others it may be considered a **spatial domain of dimension one**; e.g., a **transect along an ecosystem**, or **the midline of a river**, or a **road**.

It is of interest also to include some quantity (a **mark**) to the occurrence of the event at time t. For **earthquakes**, this quantity may be *intensity*, *magnitude*, and *energy*. For **rain events**, the quantity may be *rainfall intensity*. Associating a quantity y_i to the time t_i we have a **marked Poisson process**. We assume that the random variable describing quantity is independent from the random variable describing arrival times.

The sum of all marks for arrivals occurring in the interval t is called a **compound Poisson process**.

As an example, think about **modeling rainfall for every day of a month**. **A rainy or wet day** be decided upon a **Poisson process**, and the **mark** would be the **amount of rain for that day if it is a wet day**. The frequency distribution of rainfall in rainy days at a site determines the amount of rain, once a day is selected as wet (Richardson and Nicks, 1990). Daily rainfall distribution is skewed toward low values and it varies month to month according to climatic records.

The most typical distributions for rainfall amount are:

- **exponential** and **Weibull**,
- **gamma** and **generalised gamma**, and
- **skewed normal**, **log-normal** and **log-logistic**.

Note: In general, most of these distributions are from **generalised gamma family** or related distributions.

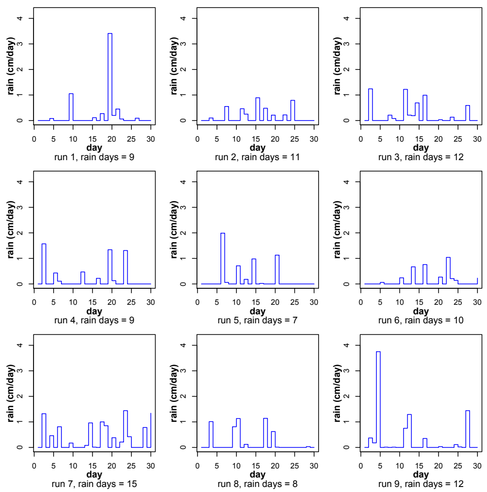


Figure: Amount of rain for a day (cm/day) during 30-day period – marked Poisson process simulation

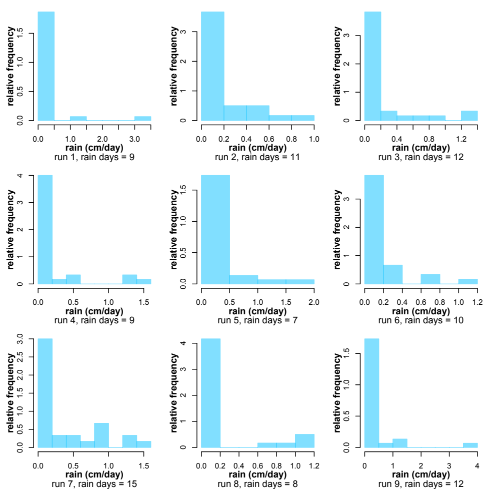


Figure: Amount of rain for a day (cm/day) during 30-day period – marked Poisson process simulation

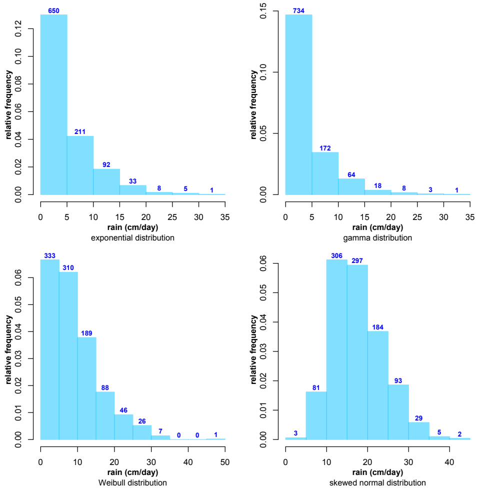


Figure: Daily rainfall amount – simulations, numer of days n = 1000