AALBORG UNIVERSITET

Exam notes

4th semester: *Probability theory*

Kasper Rosenkrands

Contents

1 Basics of probability (incl. combinatorics, law of total probability and Bayes' formula).

1.1 Axioms of Probability

Definition 1.3 (Axioms of Probability). A *probability measure* is a function P, which assigns to each event A a number of $P(A)$ satisfying

(a) $0 \leq P(A) \leq 1$

(b) P(S) = 1

(c) If $A_1, A_2, \ldots$ is a sequence of *pairwise disjoint* events, that is, if $i \neq j$, then $A_i \cap A_j = \emptyset$, then

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k)$$

1.2 Conditional Probability

Definition 1.4. Let B be an event such that $P(B) > 0$. For any event A, denote and define the *conditional probability* of A *given* B as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{or}$$
$$P(A|B)P(B) = P(A \cap B)$$

1.3 Law of Total Probability

Theorem 1.1 (Law of Total Probability). Let $B_1, B_2, \ldots$ be a sequence of events such that

(a) $P(B_k) > 0$ for $k = 1, 2, \ldots$

(b) B_i og B_j are disjoint whenever $i \neq j$

(c) $S = \bigcup_{k=1}^{\infty} B_k$

Then, for any event A, we have

$$P(A) = \sum_{k=1}^{\infty} P(A|B_k)P(B_k).$$

Proof. First note that

$$A = A \cap S = \bigcup_{k=1}^{\infty} (A \cap B_k),$$

by the distributive law for infinite unions. Since $A \cap B_1, A \cap B_2, \ldots$ are pairwise disjoint, we get

$$P(A) = P\left(\bigcup_{k=1}^{\infty} (A \cap B_k)\right) = \sum_{k=1}^{\infty} P(A \cap B_k) = \sum_{k=1}^{\infty} P(A|B_k)P(B_k).$$

Which proves the theorem. Note that the result also holds for finite sequences. ■

Corollary 1.6. If $0 < P(B) < 1$, for $B \subseteq S$, then

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

1.4 Bayes Formula

Proposition 1.11 (Bayes' Formula). Under the same assumptions as in the law of total probability and if $P(A) > 0$, then for any event B_j, we have

$$\begin{aligned} P(B_j|A) &= \frac{P(A|B_j)P(B_j)}{\sum_{k=1}^{\infty} P(A|B_k)P(B_k)} \\ &= \frac{P(A|B_j)P(B_j)}{P(A)} \qquad \text{according to the LTP.} \end{aligned}$$

Proof. Note that, by the law of total probability, the denominator is nothing but $P(A)$, and hence we must show that

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)}$$

which is to say that

$$P(B_j|A)P(A) = P(A|B_j)P(B_j)$$

which is true since both sides equal $P(A \cap B_j)$, by the definition of conditional probability. ■

Corollary 1.7. If $0 < P(B) < 1$ and $P(A) > 0$, then

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

2 Discrete stochastic variables and distributions (incl. means and variances).

2.1 Discrete Random Variable

Definition 2.1. A random variable is a real random variable that gets its values from a random experiment.

$$X : S \to \mathbb{R}.$$

Definition 2.2. If the range of X is countable, then X is called a *discrete random variable.*

2.2 Probability Mass Function

Definition 2.3. Let X be a discrete random variable with range $\{x_1, x_2, \dots\}$ (finite or countably infinite). The function

$$p(x_k) = P(X = x_k), \quad k = 1, 2, \dots$$

is called the *probability mass function* (pmf) of X.

Proposition 2.1. A function p is a possible pmf of a discrete random variable on the range $\{1, 2, \dots\}$ if and only if

(a) $p(x_k) \geq 0 \quad \text{for } k = 1, 2, \dots$

(b) $\sum_{k=1}^{\infty} p(x_k) = 1$

2.3 Cumulative Distribution Function

Definition 2.4. Let X be any random variable. The function

$$F(x) = P(X \leq x), \quad \text{for } x \in \mathbb{R},$$

is called the (*cumulative*) *distribution function* (cdf) of X.

2.4 Expected Value

Defintion 2.8. Let X be a discrete random variable with range $\{x_1, x_2, \dots\}$ (finite or countably infinite) and probability mass function p. The *expected value* of X is defined as

$$E[X] = \sum_{k=1}^{\infty} x_k p(x_k).$$

Proposition 2.12. Let X be a random variable with pmf p_X and let $g : \mathbb{R} \to \mathbb{R}$ be any function. Then

$$E[g(X)] = \sum_{k=1}^{\infty} g(x_k)p_X(x_k) \quad \text{if } X \text{ is discrete with range } \{x_1, x_2, \dots\}$$

Proposition 2.11 (Linearity for the Expectation). Let X be any random variable, and let a and b be real numbers. Then

$$E[aX + b] = aE[X] + b$$

Proof. We prove this in the discrete case, for $a > 0$. Let $Y = aX + b$, and note that Y is a discrete random variable and by definition 2.8 expected value

$$\begin{aligned} E[Y] &= \sum_{k=1}^{\infty} y_k p_Y(y_k) = \sum_{k=1}^{\infty} y_k p_X\left(\frac{y_k - b}{a}\right) \\ &= \sum_{k=1}^{\infty}(ax_k + b)p_X(x_k) \\ &= a\sum_{k=1}^{\infty} x_k p_X(x_k) + b\sum_{k=1}^{\infty} p_X(x_k) \\ &= aE[X] + b \end{aligned}$$

and we are done. ■

3 Continuous stochastic variables and distributions (incl. means and variances).

3.1 Continuous Stochastic Variable

Definition 2.5. If the cdf F is a continuous and differentiable function, then X is said to be a *continuous random variable.*

3.2 Cumulative Distribution Function

Proposition 2.3. If F is the cdf of any random variable, F has the following properties:

(a) It is nondecreasing

(b) It is right-continuous

(c) It has the limits $F(-\infty) = 0$ and $F(\infty) = 1$ (where the limits may or may not be attained at finite x).

3.3 Probability Density Function

Definition 2.6. The function $f(x) = F'(x)$ is called the *probability density function* (pdf) of X.

Proposition 2.5. Let X be a continuous random variable with pdf f and cdf F. Then

(a) $F(x) = \int_{-\infty}^{x} f(t)dt, \quad x \in \mathbb{R}$

(b) $f(x) = F'(x), \quad x \in \mathbb{R}$

(c) For $B \subseteq \mathbb{R}, \quad P(X \in B) = \int_B f(x)dx$

Proposition 2.6. A function f is a possible pdf of some continuos random variable if and only if

(a) $f(x) \geq 0, \quad x \in \mathbb{R}$

(b) $\int_{-\infty}^{\infty} f(x)dx = 1$

3.4 Proposition 2.8

Proposition 2.8. Let X be a continuous random variable with pdf X, let g be a strictly increasing or strictly decreasing, differentiable function, and let $Y = g(X)$. Then Y has pdf

$$f_Y(y) = \left|\frac{d}{dy}g^{-1}(y)\right| f_X(g^{-1}(y))$$

for y in range of Y.

Proof.

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

$$f_Y(y) = F_Y'(y) = F_X'(g^{-1}(y)) = \frac{d}{dy}g^{-1}(y) \cdot F_X'(g^{-1}(y)) = \frac{d}{dy}g^{-1}(y) \cdot f_X(g^{-1}(y))$$

■

3.5 Expected Value

Defintion 2.9. Let X be a continuous random variable with pdf f. The *expected value* of X is defined as

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

notice that the last equality is not always satisfied, but for the purpose of this course it is.

3.6 Variance

Definition 2.10. Let X be a random variable with expected value μ. The *variance* of X is defined as

$$Var[X] = E\Big[(X-\mu)^2\Big]$$

Corollary 2.2.

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

Proof. By proposition 2.12, $E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$, we have

$$\begin{aligned}
Var[X] = E\Big[(X-\mu)^2\Big] &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \\
&= \int_{-\infty}^{\infty} (x^2 - 2x\mu + \mu^2) f(x)dx \\
&= \int_{-\infty}^{\infty} x^2 f(x)dx - 2\mu \int_{-\infty}^{\infty} xf(x)dx + \mu^2 \int_{-\infty}^{\infty} f(x)dx \\
&= E[X^2] - 2\mu E[X] + \mu^2 \\
&= E[X^2] - 2E[X]^2 + (E[X])^2 \\
&= E[X^2] - (E[X])^2.
\end{aligned}$$

■

4 Two random variables: select from topics such as joint distribution, conditional distribution, independence and convolution.

4.1 Two Dimensional Random Vector

Definition 3.1. Let X and Y be random variables. The pair (X, Y) is then called a (two-dimensional) *random vector.*

4.2 Joint Cumulative Distribution Function

Definition 3.2. The *joint distribution function* (joint cdf) of (X, Y) is defined as

$$F(x, y) = P(X \leq x, Y \leq y)$$

for $x, y \in \mathbb{R}$.

Proposition 3.1 (Marginal cdf). If (X, Y) has joint cdf F, then X and Y have cdfs

$$F_X(x) = F(x, \infty) \quad \text{and} \quad F_Y(y) = F(\infty, y)$$

for $x, y \in \mathbb{R}$. Notice that there is a slight abuse of notation here, $F(x, \infty)$ refers to $\lim_{y \to \infty} F(x, y)$.

4.3 Joint Probability Density Function

Definition 3.5. If there exists a function f such that

$$P((X, Y) \in B) = \int \int_B f(x, y) dx dy$$

for all subsets $B \subseteq \mathbb{R}^2$, then X adn Y are said to be *jointly continuous.* The function f is called the *joint pdf.*

Propostion 3.3. If X and Y are jointly continuous with joint cdf F and joint pdf f, then

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y), \quad x, y \in \mathbb{R}.$$

Proposition 3.4. A function f is a possible joint pdf for the random variables X and Y if and only if

(a) $f(x, y) \geq 0$ for all $x, y \in \mathbb{R}$

(b) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx\ dy = 1$

4.4 Marginal Probability Density Function

Proposition 3.5. Suppose that X and Y are jointly continuous with joint pdf f. Then X adn Y are continuous random variables with marginal pdfs

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy, \quad x \in \mathbb{R}$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx, \quad y \in \mathbb{R}.$$

Definition 3.7. Let (X,Y) ne jointly continuous with joint pdf f. The *conditional pdf of Y given $X = x$* is defined as

$$f_Y(y|x) = \frac{f(x,y)}{f_X(x)}, \quad y \in \mathbb{R}.$$

4.5 Law of Total Probability

The following proposition is a continuous version of the law of total probability.

Proposition 3.6. Let X and Y be jointly continuous. Then

(a)

$$f_Y(y) = \int_{-\infty}^{\infty} f_Y(y|x)f_X(x)dx, \quad y \in \mathbb{R}$$

(b)

$$P(Y \in B) = \int_{-\infty}^{\infty} P(Y \in B|X = x)f_X(x)dx, \quad B \subseteq \mathbb{R}.$$

Proof. For (a), just combine Proposition 3.5 with the definition of conditional pdf for (b), part (a) gives

$$\begin{aligned} P(Y \in B) &= \int_B f_Y(y)dy = \int_B \int_{-\infty}^{\infty} f_Y(y|x)f_X(x)dx\ dy \\ &= \int_{-\infty}^{\infty} \int_B f_Y(y|x)f_X(x)dy\ dx \\ &= \int_{-\infty}^{\infty} P(Y \in B|X = x)f_X(x)dx \end{aligned}$$

as desired. ■

Proposition 3.10. Suppose that X and Y are jointly continuous with joint pdf f. Then X and Y are independent if and only if

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

for all $x, y \in \mathbb{R}$.

5 Two random variables: select from topics such as covariance and correlation, conditional expectation, conditional variance and the bivariate normal distribution.

Definition 3.10. Let y be random variable and B an event with $P(B) > 0$. The *conditional expectation* of Y given B is defined as

$$E[Y|B] = \begin{cases} \sum_{k=1}^{\infty} y_k P(Y = y_k|B) & \text{if } Y \text{ is discrete with range } \{y_1, y_2, \dots\} \\ \int_{-\infty}^{\infty} y f_Y(y|B) dy & \text{if } Y \text{ is continuous} \end{cases}$$

Definition 3.12. Suppose that X and Y are jointly continuous. We define

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f_Y(y|x) dx.$$

Following the usual intuitive interpretation, this is the expected value for Y if we know that $X = x$. The law of total expectation now takes the following form.

Propostion 3.17. Suppose that X and Y are jointly continuous. Then

$$E[Y] = \int_{-\infty}^{\infty} E[Y|X = x] f_X(x) dx.$$

Proof. By definiton of expected value and Proposition 3.6 (a)

$$E[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_Y(y|x) f_X(x) dx \; dy$$

where we cahnge the order of integration to obtain

$$E[Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_Y(y|x) dy f_X(x) dx$$

where the inner integral equals $E[Y|X = x]$ by definition, and we are done. ■

5.1 Conditional Variance

Definition 3.15. The *conditional variance* of Y given X is defined as

$$Var[Y|X] = E\Big[(Y - E[Y|X])^2 |X\Big].$$

Note that the conditonal varince is also a random variable and we think of it as the variance of Y given the value X. In particular, if we have the observed $X = x$, then we can denote and defined

$$Var[Y|X = x] = E\Big[(Y - E[Y|X = x])^2 |X = x\Big].$$

also note that if X and Y are independent, $E[Y|X] = E[Y]$, and the definition boils down to the regular variance. There is an analog of Corollary 2.2, which we leave to the reader to prove.

Corollary 3.7.

$$Var[Y|X] = E\Big[[Y^2|X]\Big] - (E[Y|X])^2$$

There is also a "law of total variance", which looks slightly more complicated than that of total expectation.

Proposition 3.19.

$$Var[Y] = Var\Big[E[Y|X]\Big] + E\Big[Var[Y|X]\Big]$$

Proof. Take expected values in Corollary 3.7 to obtain

$$E\Big[\,\mathrm{Var}[Y|X]\Big] = E[Y^2] - E\Big[(E[Y|X])^2\Big] \tag{5.1}$$

and since $E\Big[E[Y|X]\Big] = E[Y]$, we have

$$\mathrm{Var}\,\Big[E[Y|X]\Big] = E\Big[(E[Y|X])^2\Big] - (E[Y])^2 \tag{5.2}$$

and the result follows form adding (5.1) and (5.2). ∎

5.2 Covarince and Correlation

Definition 3.16. The *covariance* of X and Y is defined as

$$Cov[X,Y] = E\Big[(X - E[X])(Y - E[Y])\Big].$$

Proposition 3.20.

$$Cov[X,Y] = E[XY] - E[X]E[Y]$$

Proposition 3.21.

$$Var[X + Y = Var[x] + Var[Y] + 2Cov[X,Y].$$

Proof. By the definition of variance and covariance and repeated use of properties of expected values, we get

$$\begin{aligned} Var[X+Y] &= E\Big[(X + Y - E[X+Y])^2] \\ &= E\Big[(X - E[X] + Y - E[Y])^2] \\ &= E\Big[(X - E[X])^2 + (Y - E[Y])^2 + 2(X - E[X])(Y - E[Y])] \\ &= Var[X] + Var[Y] + Cov[X,Y] \end{aligned}$$

and we are done. ∎

Måske kan du vælge bivariate normal i stedet hvis du har tid og føler dig klog!

6 Generating functions (possibly with a focus on how probability generating functions relate to thinning of a Poisson process).

Noget af det der er med her skal postuleres uden bevis!

Proposition 3.39. Let $X_1, X_2, \ldots$ be i.i.d. nonnegative and integer valued with commen pgf G_X, and let N be nonnegative and integer valued, and independent of the X_k, with pgf G_N. Then $S_N = X_1 + \cdots + X_N$ has pgf

$$G_{S_N}(s) = G_N(G_X(s))$$

the composition of G_N and G_X.

Proof. We condition on N to obtain

$$\begin{aligned} G_{S_N}(s) &= E\left[E\left[s^{S_N}|N=n\right]\right] \\ &= \sum_{n=0}^{\infty} E[s^{S_N}|N=n]P(N=n) \\ &= \sum_{n=0}^{\infty} E[s^{S_N}]P(N=n) \end{aligned}$$

since N and S_n are independent. Now note that

$$E[s^{S_N}] = G_X(s)^n$$

by Proposition 3.38 and we get

$$G_{S_N}(s) = \sum_{n=0}^{\infty} G_X(s)^n P(N=n) = G_N(G_X(s))$$

the pgf of N evaluated at the point $G_X(s)$. ∎

6.1 The Poisson Proces

Definition 3.25. A point process where times between consecutive points are i.i.d. random variables that are $\exp(\lambda)$ is called a *Poisson process* with rate λ.

Proposition 3.41. Consider a Poisson process with rate λ, where $X(t)$ is the number of points in an interval of length t. Then

$$X(t) \sim \mathrm{Poi}(\lambda t)$$

Recall that the parameter in the Poisson distribution is also the expected value. Hence, we have

$$E[X(t)] = \lambda t$$

which makes sense since λ is the mean number of points per time unit and t is the length of the time interval. In practical applications, we need to be careful to use the same time units for λ and t.

Proposition 3.42. In a Poisson process with rate λ

(a) $T_1, T_2, \dots$ are independent and $\exp(\lambda)$

(b) $X(T_1), X(T_2), \dots$ are independent and $X(t_j) \sim Poi(\lambda t_j), \quad j = 1, 2, \dots$

6.2 Thinning and Superposition

Proposition 3.44. The thinned process is a Poisson process with rate λp.

Proof. We work with characterization (b) in Proposition 3.42. Clearly, the numbers of observed points in disjoint intervals are independent. To show the Poisson distribution, we use probability generating functions. Consider an interval of length t, letting $X(t)$ be the total number of points and $X_p(t)$ be the number of observed points in this interval. Then,

$$X_p(t) = \sum_{k=1}^{X(t)} I_k$$

where I_k is 1 if the kth point was observed and 0 otherwise. By Proposition 3.39, $X_p(t)$ has pgf

$$G_{X_p}(s) = G_{X(t)}(G_I(s))$$

where

$$G_{X(t)}(s) = e^{\lambda t(s-1)}$$

and

$$G_I(s) = 1 - p + ps$$

and we get

$$G_{X_p}(s) = e^{\lambda t(1-p+ps-1)} = e^{\lambda pt(s-1)}$$

which we recognize as the pgf of a Poisson distribution with parameter λpt. ∎

Proposition 3.45. The processes of observed and unobserved points are independent.

Proof. Fix an interval of length t, let $X(t)$ be the total number of points, and $X_p(t)$ and $X_{1-p}(t)$ the number of observed and unobserved points, respectively. Hence, $X(t) = X_p(t) + X_{1-p}(t)$ and by Proposition 3.44, we obtain

$$X_p(t) \sim \text{Poi}(\lambda pt) \quad \text{and} \quad X_{1-p}(t) \sim \text{Poi}(\lambda(1-p)t).$$

Also given that $X(t) = n$, the number of observed points has a binomial distribution with parameters n and

p. We get

$$\begin{aligned}
P(X_p(t) = j, X_{1-p}(t) = k) &= P(X_p(t) = j, X(t) = k + j) \\
&= P(X_p(t) = j | X(t) = k + j) P(X(t) = k + j) \\
&= \binom{k+j}{j} p^j (1-p)^k e^{-\lambda t} \frac{(\lambda t)^{k+j}}{(k+j)!} \\
&= \frac{(k+j)!}{k!j!} p^j (1-p)^k e^{-\lambda t} \frac{(\lambda t)^{k+j}}{(k+j)!} \\
&= e^{-\lambda p t} \frac{(\lambda p t)^j}{j!} e^{-\lambda(1-p)t} \frac{(\lambda(1-p)t)^k}{k!} \\
&= P(X_p(t) = j) P(X_{1-p}(t) = k)
\end{aligned}$$

■

7 Limit theorems.

When we introduced expected values, we argued that these could be considered averages of a large number of observations. Thus, if we have observations $X_1, X_2, \ldots, X_n$ and we do not know the mean μ, a reasonable approximation ought to be the *sample mean*

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

in other words, the average of $X_1, \ldots, X_n$. Suppose now that the X_k are i.i.d. with mean μ and variance σ^2. By the formulas for the mean and variance of sums of independent variables, we get

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} \frac{1}{n}E[X_k] = \mu$$

and

$$Var[\bar{X}] = Var\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} \frac{1}{n^2}Var[X_k] = \frac{\sigma^2}{n}$$

that is, $\bar{X}$ has the same expected value as each individual X_k and a variance that becomes smaller the larger the value of n.

7.1 The Law of Large Numbers

ALthough we can nevet guarantee that $|\bar{X} - \mu|$ is smaller than a given ε we can say that is very likely that $|\bar{X} - \mu|$ is small if n is large. That is the idea behind the following result.

Theorem 4.1. (The Law of Large Numbers). Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ, and let $\bar{X}$ be their sample mean. Then, for every $\varepsilon > 0$

$$P(|\bar{X} - \mu| > \varepsilon) \to 0 \quad \text{as} \quad n \to \infty$$

Proof. Assume that the X_k have finite variance, $\sigma^2 < \infty$. Apply Chebyshev's inequality to $\bar{X}$ and let $c = \varepsilon\sqrt{n}/\sigma$. Since $E[\bar{X}] = \mu$ and $Var[\bar{X}] = \sigma^2/n$, we get

$$P(|\bar{X} - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0 \quad \text{as} \quad n \to \infty$$

The assumptions of finite variance is neccessary for this proof to work. However, the law of large numbers is tru also if the varinace is infinite, but the proof in that case is more involved and we will not give it. ■

We say that $\bar{X}$ *converges in probability* to μ and write

$$\bar{X} \xrightarrow{P} \mu \quad \text{as} \quad n \to \infty$$

Corollary 4.1. Consider an experiment where the event A occurs with probability p. Repeat the experiment independently, let S_n be the number of times we get the event A in n trials, and let $f_n = S_n/n$, the relative frequency. Then

$$f_n \xrightarrow{P} p \quad \text{as} \quad n \to \infty$$

Proof. Define the indicators

$$I_k = \begin{cases} 1 & \text{if we get } A \text{ in the } k\text{th trial} \\ 0 & \text{otherwise} \end{cases}$$

for $k = 1, 2, \ldots, n$. Then the I_k are i.i.d. and we know from Section 2.5.1 that they have mean $\mu = p$. Since f_n is the sample mean of the I_k, the law of large numbers gives $f_n \xrightarrow{P} p$ as $n \to \infty$. ■

Theorem 4.2 (The Central Limit Theorem). Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$ and let $S_n = \sum_{k=1}^{n} X_k$. Then, for each $x \in \mathbb{R}$, we have

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \quad \xrightarrow[n\to\infty]{d} \quad N(0, 1)$$

as $n \to \infty$, where Φ is the cdf of the standard normal distribution.

Example 4.5. Consider again bufons needle problem. Recall that the probability that the randomly tossed needle intersects a line is $2/\pi$ and how we argued in the previous section that $2/f_n \xrightarrow{P} \pi$ as $n \to \infty$. Let $\hat{\pi}$ denote our estimate of π after n tosses:

$$\hat{\pi} = \frac{2}{f_n}$$

As mentioned previously, Buffon himself actually that this experiment be used to estimate π. Let us say that on some occasion he tossed the needle 1000 times. What is the probability that he got the estimate correct to two decimals?

We wish to find

$$P(|\hat{\pi} - 3.14| \le 0.005) = P\left(\frac{2}{3.145} \le f_n \le \frac{2}{3.135}\right)$$

where, by the following equation with $p = 2/\pi$

$$f_n \overset{d}{\approx} N\left(p, \frac{p(1-p)}{n}\right) \qquad f_n \overset{d}{\approx} N\left(\frac{2}{\pi}, \frac{2(\pi-2)}{n\pi^2}\right)$$

With $n = 1000$ we now get

$$\begin{aligned} P(|\hat{\pi} - 3.14| \le 0.005) &= P\left(\frac{2}{3.145} \le f_{1000} \le \frac{2}{3.135}\right) \\ &= \Phi\left(\frac{2/3.135 - 2/\pi}{\sqrt{2(\pi-2)/1000\pi^2}}\right) - \Phi\left(\frac{2/3.145 - 2/\pi}{\sqrt{2(\pi-2)/1000\pi^2}}\right) \\ &= \Phi(0.09) - \Phi(-0.05) \approx 0.06 \end{aligned}$$

i.e. the probability of being correct within two decimals after 1000 throws is 6 percent. That is no a very good reward for all that needle tossing.

8 Markov chains.

8.1 Discrete-Time Markov Chains

Definition 8.1.. Let $X_0, X_1, X_2, \ldots$ be a sequence of discrete random variables, taking values in some set S and that are such that

$$P(X_{n+1} = j | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i) = P(X_{n+1} = j | X_n = i) \tag{8.1}$$

for all $i, j, i_0, \ldots, i_{n-1}$ in S and all n. The sequence $\{X_n\}$ is then called a *Markov chain*. Furthermore (8.1) is called the *Markov property*.

We ofthen think of the index n as discrete time and say that X_n is the *state* of the chain at time n, where the state space S may be finite or coutably infinte.

In general, the probability $P(X_{n+1} = j | X_n = i)$ depends in i, j, and n. It is, however, often the case that there is no dependence on n. We call such chains *time-homogenous* and restrict our attention to these chains. Since the conditional probability in the definition thus depends only on i and j, we use the notation

$$p_{ij} = P(X_{n+1} = j | X_n = i), \quad i, j \in S$$

and call these the *transition probabilities* of the Markov chain. Thus, if the chain is in state i, the probabilities p_{ij} desribe how the chain chooses which state to jump to necxt. Obviously, the transition probabilities have to satisfy the following two criteria:

$$\text{(a) } p_{ij} \geq 0 \quad \text{for all } i, j \in S, \qquad \text{(b) } \sum_{j \in S} p_{ij} = 1 \quad \text{for all } i \in S$$

8.1.1 Classification of States

Definition 8.2. If $p_{ij}^{(n)} > 0$ for some n, we say that state j is *accesible* from state i, written $i \to j$. If $j \to i$, we say that i and j *communicate* and write this $i \leftrightarrow j$.

In general, if we fix a state i in the state space of a Mrkov vhain, we can find all states that communicate with i and form a *communicating class* containing i. It is easy to realize that not only does i communicate with all states in this class but they all communicate with each other. By convention, every state communicates with itself (it can reach itself in 0 steps) so every state belongs to a class. If you wish to be more mathemaical, the relastion "$\leftrightarrow$" is an equivalence relation and thus divides the state space into equivalence classes that are precisely the communicating classes.

Definition 8.3. If all states in S communicate with each other, the Markov chain is said to be *irreducible*.

Definition 8.4. Consider a state $i \in S$ and the τ_i be the number of steps it takes for the chain to first visit i. Thus

$$\tau_i = \min\{n \geq 1 \ : \ X_n = i\}$$

where $\tau_i = \infty$ if i is never visited. If $P_i(\tau_i < \infty) = 1$, the state i is said to be *recurrent* and if $P_i(\tau_i < \infty) < 1$, it is said to be *transient*.

Proposition 8.1. State i is

$$\text{transient if} \quad \sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$$
$$\text{recurrent if} \quad \sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$$

8.1.2 Stationary Distribution

Definition 8.5. Let P be the transition matrix of a Markov chain with state space S. A probability distribution $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ on S satisfying

$$\boldsymbol{\pi} P = \boldsymbol{\pi}$$

is called a *stationary distribution* of the chain.

Proposition 8.2. Consider an irreducible Markov chain. If a stationary distribution exists, it is unique.

Proposition 8.3. If S is finite and the Markov chain is irreducible, a unique stationary distribution $\boldsymbol{\pi}$ exists.

Definition 8.6. Let i be a recurrent state. If $E_i[\tau_i] < \infty$, then i is said to be *positive recurrent.* If $E_i[\tau_i] = \infty$, i is said to be *null recurrent.*

Corollary 8.4. For an irreducible Markov chain, there are three possibilities: **(a)** all states are positive recurrent, **(b)** all states are null recurrent, and **(c)** all states are transient.

8.1.3 Convergence to the Stationary Distribution

Definition 8.8. The *period* of state i is defined as

$$d(i) = \gcd\{n \geq 1 \ : \ p_{ii}^{(n)} > 0\}$$

the greatest common divisor of lengths of cycles through which it is possible to return to i. If $d(i) = 1$, state i is said to be *aperiodic*; otherwise it is called *periodic.* Note that this is a class property, which means if $i \leftrightarrow j$ and either i or j is periodic so is the other. Likewise for aperiodic.

Theorem 8.1. Consider an irreducible, positive recurrent, and aperiodic Markov chain with stationary distribution $\boldsymbol{\pi}$ and n-step transition probabilities $p_{ij}^{(n)}$. Then

$$p_{ij}^{(n)} \to \pi_j \quad \text{as} \quad n \to \infty$$

for all $i, j \in S$.

An irreducible, positve recurrent, and aperiodic Markov chain is called *ergodic.*

9 Stochastic simulation.

9.1 Simulation of Continuous Distributions

Proposition 5.2 (The Inverse Transformation Method). Let F be a distribution function that is continuous and strictly increasing. Further, let $U \sim \text{unif}[0,1]$ and define the random variable $Y = F^{-1}(U)$. Then Y has distribution function F.

Proof. Start with F_Y, the distribution function of Y. Take x in the range of Y to obtain

$$F_Y(x) = P(F^{-1}(U) \leq x) \qquad = P(U \leq F(x)) = F_U(F(x)) = F(x)$$

where the last equality follows since $F_U(u) = u$ if $0 \leq u \leq 1$. The argument here is $u = F(x)$, which is between 0 and 1 since F is a cdf. ■

Example 5.2. Generate an observation from an exponential distribution with parameter λ.

Here,

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0$$

To find the inverse, as usual solve $F(x) = u$, to obtain

$$x = F^{-1}(u) = -\frac{1}{\lambda}\log(1-u), \quad 0 \leq u < 1$$

Hence if $U \sim \text{unif}[0,1]$, the random variable

$$X = -\frac{1}{\lambda}\log(1-U)$$

is $\exp(\lambda)$. We can note here that since $1-U$ is also uniform on $[0,1]$, we might as well take $X = -\log U/\lambda$.

Proposition 5.3 (The Rejection Method).

1. Generate Y and $U \sim \text{unif}[0,1]$ independent of each other.
2. If $U \leq \frac{f(Y)}{cg(Y)}$, set $X = Y$. Otherwise return to step 1.

The random variable X generated by the algorithm has pdf f.

Proof. Let us first make sure that the algorithm terminates. The probability in any given step 2 to accept Y is, by Corollary 3.2,

$$\begin{aligned} P\left(U \leq \frac{f(Y)}{cg(y)}\right) &= \int_{\mathbb{R}} P\left(U \leq \frac{f(Y)}{cg(y)}\right) g(y)dy \\ &= \int_{\mathbb{R}} \frac{f(Y)}{cg(y)} g(y)dy = \frac{1}{c}\int_{\mathbb{R}} f(y)dy = \frac{1}{c} \end{aligned}$$

where we used the independence of U and Y and the fact that $U \sim \text{unif}[0,1]$. Hence, the number of iterations until we accept a value has a geometric distribution with success probability $1/c$. The algorithm therefore always terminates in a number of steps with mean c from which it also follows that we should choose c as small as possible.

Next we turn to the question of why this gives the corresct distribution. To show this, we will show that the consitional distribution of Y, given acceptance, is the same as the distribution of X. Recalling the definition of conditional probability and the fact that the probability of acceptance is $1/c$, we get

$$P\left(Y \leq x \;\middle|\; U \leq \frac{f(Y)}{cg(y)}\right) = cP\left(Y \leq x \,\cap\, U \leq \frac{f(Y)}{cg(y)}\right)$$

By independence, the joint pdf of (U, Y) is $f(u, y) = g(y)$, and the above expression becomes

$$\begin{aligned} cP\left(Y \leq x \,\cap\, U \leq \frac{f(Y)}{cg(y)}\right) &= c\int_{-\infty}^{x}\int_{0}^{f(y)/cg(y)} g(y)du\ dy \\ &= c\int_{-\infty}^{x} \frac{f(y)}{cg(y)} g(y)dy = P(X \leq x) \end{aligned}$$

which is what we wanted to prove. ■