Introduction

My notes from the legendary book of Papoulis and Pillai [1]. The book assumes that the reader does already have fairly strong background in calculus and linear algebra. I used two books [2, 3] as assistant for evaluating some of the mathematical identities or integrals.

This book made me realize that a good technical book is not necessarily one that's easy to follow – one that shows you very easily how each identity or result is arrived at. Some of the results in the book are presented without much explanation, and this required me to show a lot of effort to understand them; this effort seems to lead to a staying power.

Chapter 1

The meaning of probability

Probabilistic interpretations are [...] necessary because of our ignorance.

1.1 Notations and basic definitions

- The *certain event* S is the event that occurs in every trial.
- The union of two events A, B is denoted with $A \cup B$ or $A + B$.
- The intersection of two events A, B is denoted with $A \cap B = AB$.
- Two sets A, B are *mutually exclusive* if they have no common elements, *i.e.* $AB = \emptyset$.
- *Partition* is a collection mutually exclusive subsets $A_1, \ldots, A_n$ whose union equals S, *i.e.* $\bigcup_{i=1}^{n} A_i = S$.

Chapter 2

The Axioms of Probability

2.1 The Axioms

We assign to each event A a number $P(A)$, which we call *the probability of the event* A. This number is so chosen to satisfy three conditions:

I. $P(A) \geq 0$

II. $P(S) = 1$

III. if $AB = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.

2.2 Definitions and some basic laws

- De Morgan's Law: For two sets A and B,
$$\overline{A \cup B} = \overline{A}\ \overline{B}$$
$$\overline{AB} = \overline{A} \cup \overline{B}$$
- Two events A, B are *equal with probability 1* (*i.e.* both contain the same elements) iff $P(A) = P(B) = P(AB)$
- *Events* are subsets of S to which we have assigned probabilities.
- The class $\mathbf{F}$ of events is a nonempty class of sets such that if A, B are events, then $A + B$ and AB are also events:
$$\text{If } A \in \mathbf{F} \text{ then } \overline{A} \in \mathbf{F}$$
$$\text{If } A \in \mathbf{F} \text{ and } B \in \mathbf{F} \text{ then } A + B \in \mathbf{F}$$
- *Field.* The two properties above give a minimum set of conditions for $\mathbf{F}$ to be a field.
- *Borel fields.* Suppose $A_1, \ldots, A_n, \ldots$ is an infinite sequence of sets in $\mathbf{F}$. If the union and intersection of these sets also belongs to $\mathbf{F}$ then $\mathbf{F}$ is called a Borel field.
- *Axiomatic definition of an experiment.* In theory of probability, an experiment is specified in terms of the following concepts:
  1. The set S of all experimental outcomes
  2. The Borel field of all events of S.
  3. The probabilities of these events.

  Example. The (fair) die experiment has 6 experimental outcomes. Two of the events are *even, odd*; each of those events is a subset of three experimental outcomes with probability 0.5.
- *Conditional probability* of an event A conditioned on another event M is
$$P(A|M) = \frac{P(AM)}{P(M)}$$
- **Fundamental remark.** Conditional probabilities satisfy the three probability axioms:

I. $P(A|M) = \frac{P(AM)}{P(M)}$

II. $P(S|M) = 1$

III. If A, B are mutually exclusive, then $P(A \cup B|M) = P(A|M) + P(B|M)$

- *Total probability theorem.* If $\mathbf{U} = [A_1, \ldots, A_n]$ is a partition of S and B is an arbitrary event, then

$$P(B) = P(B|A_1)P(A_1) + \ldots + P(B|A_n)P(A_n)$$

.

- *Bayes' theorem* follows from the theorem above and the fact that $P(BA_i) = P(B|A_i)P(A_i) = P(A_i|B)P(B)$:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + \ldots P(B|A_n)P(A_n)}$$

- *Independence.* The events A, B are independent if $P(AB) = P(A)P(B)$

- *Mutual independence.* The events A_k for $k \in K$ are independent iff $P(\bigcap_{m\in M}) = \bigcap_{m\in M} P(A_m)$ for any $M \subset K$.

  Example. For the four events A, B, C, D to be independent, we must have:
  $P(ABCD) = P(A)P(B)P(C)P(D)$,
  $P(ABC) = P(A)P(B)P(C)$, $P(ABD) = P(A)P(B)P(D)$,
  $P(ACD) = P(A)P(C)P(D)$, $P(BCD) = P(B)P(C)P(D)$,
  $P(AB) = P(A)P(B)$, $P(AC) = P(A)P(C)$, $P(AD) = P(A)P(D)$,
  $P(BC) = P(B)P(C)$, $P(BD) = P(B)P(D)$, $P(CD) = P(C)P(D)$.

  Example. It is possible for events to be pairwise independent but not mutually independent.

Chapter 3

Repeated Trials

Combining two experiments, such as tossing a coin and rolling a die, requires the definition of a new set S of all experimental outcomes. This set is constructed via Cartesian product:

$$S = S_1 \times S_,2$$

where S_1 (S_2) is set of experimental outcomes the first (second) experiment.

The same can be applied for repeated experiments. If we toss a coin n times, the set of experimental outcomes becomes

$$S = S_1 \times \ldots \times S_n.$$

3.1 Bernoulli Trials

- *Permutation* is a distinct arrangement of n objects; *e.g. bac* is a permutation of *a,b,c.*

  Permutation applies also to $k < n$ objects. For example, *ba* is a permutation of $k = 2$ objects in *a,b,c.*

- The total number of permutations of n objects taken k at a time is $\frac{n!}{(n-k)!}$; *e.g.* , *ab, ac, bc, ba, ca, cb* are the permutations of *a,b,c* for $k = 1$.

- The number above pays attention to distinct orderings; *i.e.* two permutations that have the same objects in different orders count as two.

- *Combination* does not pay attention to the number of orderings. So the total combinations of n objects taken k at a time is $\binom{n}{k} := \frac{n!}{(n-k)!k!}$. *E.g., ab, ac, bc* are the combinations of *a,b,c* for $k = 2$.

- **Fundamental theorem.** Let us repeat a binary experiment n times. Then,

$$p_n(k) := P\{A \text{ occurs } k \text{ times in any order}\} = \binom{n}{k} p^k q^{n-k}$$

- **Bernoulli's Theorem** is fundamental as it serves as a bridge between the axiomatic and frequency definitions of probability. In other words, it provides justification for the relative frequency interpretation.

  Let A be an event whose probability of occurrence in a single trial is p. If k denotes the number of occurrences of A in n independent trials, then

$$P\left(\left|\frac{k}{n} - p\right| > \epsilon\right) < \frac{pq}{n\epsilon^2}.$$

  In other words, the probability that the frequency interpretation is *not* justified can be made arbitrarily small [*i.e.* $P(|k/n - p| > \epsilon)$ for any $\epsilon > 0$] provided that the number of trial is above a certain limit.

Chapter 4

The Concept of a Random Variable

4.1 Definitions and some basic properties

- **Definition of Random Variable.** An RV $\mathbf{x}$ is a process of assigning a number $\mathbf{x}(\zeta)$ to every outcome ζ.

  Example 1. In the coin tossing experiment, if $\zeta_1 = head$ and $\zeta_2 = tails$, one may construct the RV $\mathbf{x}$ such that $\mathbf{x}(\zeta_1) = 0$ and $\mathbf{x}(\zeta_2) = 1$.

  Example 2. Suppose that our RV $\mathbf{x}$ will measure temperature. Then, one can construct the RV $\mathbf{x}$ simply as $\mathbf{x}(\zeta_i) = \zeta_i$ (*e.g.* when ζ is a continuous quantity such as temperature). In this case, the variable ζ has a double meaning (see p77): It is the outcome of the experiment and the corresponding value $\mathbf{x}(t)$ of the RV $\mathbf{x}$.

  The construction of RV is important for accurately constructing the intended probability distribution. For example, in the experiment of tossing a coin twice, if we don't care about the order of the outcomes, one can construct the RV $\mathbf{x}$ as:

  $$\mathbf{x}(HH) = 0,\ \ \mathbf{x}(HT) = 1,\ \ \mathbf{x}(TT) = 2.$$

  But if we do care about the order, then we can construct $\mathbf{x}$ as:

  $$\mathbf{x}(HH) = 0,\ \ \mathbf{x}(HT) = 1,\ \ \mathbf{x}(TH) = 2,\ \ \mathbf{x}(TT) = 3.$$

- **Probability distribution function.** The probability $P\{\mathbf{x} \le x\}$ of the event $\{\mathbf{x} \le x\}$ is function of x, called the *(cumulative) distribution function* of the RV $\mathbf{x}$. Some properties of distribution functions are:
  - $F(+\infty) = 1$ and $F(-\infty) = 0$.
  - It is a nondecreasing function of x; *i.e.* $x_1 < x_2 \implies F(x_1) \le F(x_2)$.
  - If $F(x_0) = 0$, then $F(x) = 0$ for every $x \le x_0$.
  - $P\{\mathbf{x} > x\} = 1 - F(x)$.
  - The function $F(x)$ is continuous from the right $F(x^+) = \lim_{\epsilon\to 0} F(x+\epsilon) = F(x)$
  - $P\{x_1 < \mathbf{x} \le x_2\} = F(x_2) - F(x_1)$.
  - $P\{\mathbf{x} = x\} = F(x) - F(x^-)$ where $F(x^-) = \lim_{\epsilon\to 0} F(x-\epsilon)$
  - $P\{x_1 \le \mathbf{x} \le x_2\} = F(x_2) - F(x_1)$.
- **Probability density function.** The derivative of the probability distribution function $F_x(x)$ is called the probability density function of $\mathbf{x}$. Thus

  $$f_x(x) := \frac{dF_x(x)}{dx} = \lim_{\Delta x\to 0} \frac{F_x(x+\Delta x) - F_x(x)}{\Delta x} \ge 0.$$

- **Conditional Distributions**. The conditional distribution $F(x|M)$ of an RV $\mathbf{x}$ assuming M is:

  $$F(x|M) = P\{\mathbf{x} \le x|M\} = \frac{P\{\mathbf{x} \le x, M\}}{P(M)},$$

  where $P\{\mathbf{x} \le x, M\}$ is the intersection of the events $\{\mathbf{x} \le x\}$ and M; that is, the event consisting of all outcomes ζ such that $\mathbf{x}(\zeta) \le x$ and $\zeta \in M$.
  - Suppose that $M = \{\mathbf{x} \le a\}$. Then,

    $$F(x|M) = \begin{cases} 1 & \text{if } x \ge a \\ \frac{F(x)}{F(a)} & \text{if } x < a \end{cases}$$

    Moreover,

    $$f(x|M) = \begin{cases} \frac{f(x)}{F(a)} & \text{if } x < a \\ 0 & \text{otherwise} \end{cases}$$

– Suppose that $M = \{b < \mathbf{x} \leq a\}$. In this case,

$$F(x|M) = \begin{cases} 1 & x > a \\ \frac{F(x)-F(b)}{F(a)-F(b)} & b \leq x < a \\ 0 & x < b \end{cases}$$

The corresponding Density is:

$$f(x|M) = \begin{cases} \frac{f(x)}{F(a)-F(b)} & \text{if } x < a \\ 0 & \text{otherwise} \end{cases}$$

4.2 Some Continuous Random Variables

- **Normal (Gaussian) Distribution**. Its density function is

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2)/2\sigma^2}.$$

Its distribution function is $F_x(x) = \int_{-\infty}^{x} f_x(x) := G(\frac{x-\mu}{\sigma})$ where the function $G(x)$

$$G(x) := \int_{-\infty}^{x} e^{-y^2/2} dy$$

is often available in tabulated form and is directly related to the erf function; $\int_0^x e^{-y^2/2} dy = G(x) - \frac{1}{2}$.

- **Exponential distribution.**

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 \text{otherwise} \end{cases}$$

. The exponential distribution has the *memoryless* propert; that is, for $s, t \geq 0$, $P\{\mathbf{x} > t + s | \mathbf{x} > s\} = P\{\mathbf{x} > t\}$. In other words, the time that passed, s, doesn't change anything for the exponential distro.

4.3 Some Discrete Random Variables

- **Binomial distribution.** $\mathbf{y}$ is said to be a Binomial RV with parameters n and p if $\mathbf{y}$ takes values $0, 1, \ldots, n$ with

$$P\{\mathbf{y} = k\} = \binom{n}{k} p^k q^{n-k} \qquad p + q = 1 \qquad k = 0, 1, \ldots, n$$

- **Poisson distribution** is similar to Binomial; it represents the number of occurrences of a *rare* events in a large number of trials.

$$P\{\mathbf{x} = k\} = \lambda^k \frac{e^{-\lambda}}{k!} \quad k = 0, 1, 2, \ldots, \infty$$

- **Geometric distribution** represents the probability of first success against the number of trials.

$$P\{\mathbf{x} = k\} = P(\underbrace{\overline{A}\overline{A}\ldots\overline{A}}_{k-1} A) = P(\overline{A})P(\overline{A})\ldots P(\overline{A})P(A) = (1-p)^{k-1}p = pq^{k-1}$$

- **Negative binomial distribution** generalizes geometric distribution; it's used to answer the question "how many trials are needed for r" successes.

$$P\{y = k\} = \binom{k-1}{r-1} p^r q^{k-r}.$$

4.4 Approximations to Binomial Distribution

Despite its enormous importance, computing the binomial distribution can be involved particularly for large n. There are two widespread approximations of the binomial for two different situations.

- **The Normal Approximation (DeMoivre-Laplace Theorem)**. This is used when k is in the $\sqrt{npq}$ neighbourhood of np; that is, when k is near the peak of the distribution and p is not too close to 0 or 1. In such cases;

$$\binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi npq}} e^{-(k-np)^2/2npq} \qquad p+q=1$$

and (with error correction — see p108)

$$P\{k_1 \leq \mathbf{x} \leq k_2\} \approx G\left(\frac{k_2+0.5-np}{\sqrt{npq}}\right) - G\left(\frac{k_1-0.5-np}{\sqrt{npq}}\right).$$

Generalization to m events.

$$\frac{n!}{k_1!\ldots k_m!} p_1^{k_1}\ldots p_m^{k_m} \approx \frac{\exp\{-\frac{1}{2}\left[\frac{(k_1-np_1)^2}{np_1}+\ldots\frac{(k_r-np_r)^2}{np_r}\right]\}}{\sqrt{(2\pi n)^{r-1}p_1\ldots p_r}}$$

- **The Poisson Approximation.** The Gaussian approximation deteriorates one of the two events happens very rarely; *i.e.* as $p \to 0$ or $p \to 1$. However, there are many events that fall in this category (*e.g.* number of calls on a telephone line, claims in an insurance company etc). In this case,

$$\binom{n}{k} p^k q^{n-k} \approx e^{-\lambda}\frac{\lambda^k}{k!}$$

and

$$P\{k_1 \leq \mathbf{x} \leq k_2\} \approx e^{-np}\sum_{k=k_1}^{k_2}\frac{(np)^k}{k!}$$

.

Generalization to $m+1$ events. Suppose that $A_1,\ldots,A_{m+1}$ are the $m+1$ events of a partition with $PA_i = p_i$. If $np_i \to a_i$ for $i \leq m$ (note that the $(m+1)$th event is excluded), then

$$\frac{n!}{k_1!\ldots k_{m+1}!} p_1^{k_1}\ldots p_{m+1}^{k_{m+1}} \approx \frac{e^{-a_1}a_1^{k_1}}{k_1!}\ldots\frac{e^{-a_m}a_m^{k_m}}{k_m!}$$

Chapter 5

Sequences of Random Variables

5.1 Notes

- Multivariate Transformation
- How to integrate some RVs from a multi-variate distro to obtain ... (Sec 7.2)
- How to compute the mean conditioned on a subset of RVs
- Characteristic function leads to so much interesting applications, such as:
- Computing the PDF of Bernoulli from Binomials (#256)
- Computing the PDF of Poisson from Binomials (#256)
- The sum of jointly normal RVs is normal (#257)
- The sum of the squares of independent normal variables is chi-square (#259)
- The sum of two chi-square distros is also chi-square (#260)
- The optimal single-value estimation (in the MS sense), c, of a future value of a RV $\mathbf{y}$ is $c = E(\mathbf{y})$.
- The optimal functional estimation of $\mathbf{y}$ *i.t.o.* a (dependent) RV $\mathbf{x}$ is $c(x) = E(\mathbf{y}|x)$. In general, this estimation is *non-linear*.
- The optimal linear f. estimation of $\mathbf{y}$ *i.t.o.* $\mathbf{x}$ is $c(x) = Ax + B$ where $B = \eta_y - A\eta_x$ and $A = r\sigma_x/\sigma_y$.
- *For Gaussian RVs, linear and non-linear MS estimators are identical (#264).*
- *The orthogonality principle*: The error between of linear estimator $\hat{\mathbf{y}} = A\mathbf{x} + B$ of $\mathbf{y}$ is orthogonal to the data: $E\{(\mathbf{y} - \hat{\mathbf{y}})\mathbf{x}\} = 0$
- Generalizes to the linear estimate of $\mathbf{s}$ *i.t.o.* multi RVs $\mathbf{x}_1, \ldots, \mathbf{x}_n$: $E\{(\mathbf{s} - \hat{\mathbf{s}})\mathbf{x}_i\}$ for $i = 1, \ldots, n$.
- Generalizes to *non-linear* estimation: $E\{[\mathbf{s} - g(\mathbf{X})]w(\mathbf{X})\}$, where $g(\mathbf{X})$ is the non-linear MS estimator and $w(\mathbf{X})$ *any* function of data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. (#269)
- Computing the linear MS estimator is *much easier* if the RVs $\mathbf{x}_i$ are orthogonal to one another (*i.e.* $R_{ij} = 0$ for $i \neq j$). That is why it is often we perform *whitening* (see #271-272).
- Convergence modes for a random sequence (RS) $\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots$
  - Everywhere (e), almost everywhere (a.e.), in the MS sense (MS), in prob (p), in distro (d).
  - Ordered in relaxedness (except last two): d > p > (a.e. | MS)
  - Cauchy criterion (CC): We typically think of convergence as converging into a sequence x. With CC, we can eliminate this need. That is, we ask that $|x_{n+m} - x_n| \to 0$ as $n \to \infty$.
- All the below are technically applications of RS convergence:
  - The law of large numbers: If p is the prob of event A in a single experiment and k is number of successes in n trials, then we can show that p tends to k/n in *probability*.
  - Strong law of large numbers: We can even show that p tends to k/n almost everywhere (but proof is more complicated).

- The ones below are more specifically applications for estimating $E\{\bar{\mathbf{x}}_n\}$ (and optionally $\bar{\sigma}_n^2$): the sample mean of a RS of n RVs $\bar{\mathbf{x}}_n = \frac{\mathbf{x}_1+\ldots+\mathbf{x}_n}{n}$ (the variance of the sample mean)
  - Markov's thm: If the RVs $\mathbf{x}_i$ are s.t. the mean of $\bar{\eta}_n$ of $\bar{\mathbf{x}}_n$ tends to a limit η and its variance $\bar{\sigma}_n$ tends to 0 as $n \to \infty$, then $E\{(\bar{\mathbf{x}}_n - \eta)^2\} \to 0$ as $n \to \infty$. Convergence in MS sense.
    (Note that $\mathbf{x}_i$ do not have to be uncorrelated or independent)
  - Corollary: if $\mathbf{x}_i$ are uncorrelated and $\frac{\sigma_1^2+\ldots+\sigma_n^2}{n} \to 0$ as $n \to \infty$, then $\bar{\mathbf{x}}_n \to \eta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E\{\mathbf{x}_n\}$ as $n \to \infty$. Convergence in MS sense.
    (Note that now we do require uncorrelated RVs but in exchange, compared to Markov's thm, the condition on the mean is removed and we *do* have a way of computing the mean η.)
  - Khinchin's thm: the above two required us to know *something* about the variance. According to Khinchin, if $\mathbf{x}_i$ are i.i.d. (stricter condition), then we $\bar{\mathbf{x}}_n$ tends to η even if we know nothing about the variance of $\mathbf{x}_i$'s. However, now we have convergence in probability only.
- The Central Limit Theorem (CLT) is also application of RS conv.—conv. in *distribution*:
  - Given n independent RVs $\mathbf{x}_i$, we form their sum (not sample mean!) $\mathbf{x} = \mathbf{x}_1 + \ldots + \mathbf{x}_n$. This is an RV with mean η and σ. CLT states that the distro $F(x)$ of $\mathbf{x}$ approaches a *normal distro* with the same mean and variance: $F(x) \approx G(\frac{x-\eta}{\sigma})$.
  - If the RVs are continuous, then $f(x)$, the *density* of x, also approaches a normal density.
  - The approximations become *exact* asymptotically (*i.e.* as $n \to \infty$).
  - Good n values: if $\mathbf{x}_i$ are i.i.d. , then $n = 30$ is adequate for most applications. If the functions $f_i(x)$ are smooth, even $n = 5$ can be enough.
  - Error corr: The errors of CLT can be corrected by matching higher order moments using Hermite polynomials (see#281 and also Section 6.8.1).
  - CLT for products $\mathbf{y} = \mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_n$ (assuming $\mathbf{x}_i > 0$) is *lognormal*. (Easily seen by letting $\mathbf{z} = \ln \mathbf{y} = \ln \mathbf{x}_1 + \ldots + \ln \mathbf{x}_n$ and $\mathbf{y} = e^{\mathbf{z}}$.
- Sampling random sequences for a target distribution $F(x)$
- The general idea is to generate a uniform RS $u_1, \ldots, u_n$, and then use some method to convert it to a RS $x_1, \ldots, x_{n'}$ (where $n' \leq n$) that matches the distribution of $F(x)$.
- There are several methods for sampling
  - Percentile Transformation Method: This is a straightforward method ($x_i = F_x^{-1}(u_i)$) but we can apply it only if we can compute the inverse of the target distro, $F_x^{-1}(u)$. We can use this also to transform a sequence y_i with a non-uniform distribution: $x_i = F_x^{-1}(F_y(y_i))$ (see #292).
  - Rejection method: Allows sampling without F_x^{-1}. The idea is to find an event M such that $f_x(x|M) = f_y(x)$. Turns out that this event is $M = \{\mathbf{u} \leq r(\mathbf{x})\}$ where $r(\mathbf{x}) = af_y(x)/f_x(x)$. To implement this, one needs to find the upper bound (i.e. a; see Example 7-23).
  - Mixing method: $f_x(x) = p_1 f_1(x) + \ldots + p_m f_m(x)$. See proof in #294 for explanation.
  - Some transformations:
    - Binomial RV: Sum of m i.i.d. Bernoulli's equals Binomial (#294).
    - Erlang: Sum of m i.i.d. exponentials equals Erlang
    - Chi-square: Sum an appropriate Erlang and a Normal
    - Student-t: Use a normal and a chi-square
    - Log-normal: Use a normal
  - Generating Normal RSs
    - Clearly, we can use CLT and sum some uniform RSs
    - Rejection and mixing: We can mix two Truncated normal RVs obtained via rejection
    - Polar coordinates: Use a Rayleigh and a uniform to obtain two independent normals
    - Use two uniforms to obtain two normals (see Example 6-15 to follow the equation right after 7-172)

5.2 Interesting identities/lemmas/theorems

- The unbiased linear estimator with minimum variance is the one shown in (7-17).
- Theorem 7.1: The correlation matrix is nonnegative definite
- Sum of jointly normal RVs is normal
- Sample mean and variance of n RVs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n}$ and $\bar{\mathbf{v}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x} - \bar{\mathbf{x}})^2$.
- Goodman's theorem: The statistics of a real zero-mean n-dimensional normal RV are completely determined i.t.o. the n^2 parameters—the elements of its covariance matrix. However, a similar *complex* RV requires $2n^2 + n$ elements. Goodman's theorem (#259) gives a special class of normal complex RVs that are determined completely with n^2 parameters only.
- Berry-Esseen Theorem: Gives an upper bound for the approx. error of CLT, given that the third order of the unknown distro is finite and the variables $\mathbf{x}_i$ are i.i.d.

5.3 Important concepts

- Group independence
- Correlation and covariance matrices
- The orthogonality principle
- Mean square estimation (Linear and non-linear)

5.4 Some terminology

- Homogeneous linear estimation: Linear estimation without a bias term
- Nonhomogeneous linear estimation: Linear estimation with a bias term

5.5 Redo in future

- Poisson
- Chi square
- Show why sample variance is divided by $n - 1$.
- Prove why the generalized orthogonality principle (7-92) leads to optimal MS estimators (linear or non-linear).
- Order convergence modes
- Prove the law of large numbers

Chapter 6

Statistics

6.1 Definitions

- *Statistic.* Any function of the sample RV vector, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, is called a *statistic*. (This definition applies only for this chapter! See footnote in p306)
- *Unbiased estimator.* An estimator $T(\mathbf{X})$ is an unbiased estimator for θ if $E\{T(\mathbf{X})\} = \theta$.
- *Point estimator* is an RV $\hat{\boldsymbol{\theta}} = g(\mathbf{X})$ that estimates a parameter θ.
- *Point estimate* $\hat{\theta}$ is a function of the *observation vector* $X = [x_1, \ldots, x_n]$ and the corresponding RV; *i.e.* $\hat{\theta} = g(X)$
- *Consistent estimator.* Let $\hat{\boldsymbol{\theta}}_n = g_n(\mathbf{X})$ be estimator of parameter θ based on a sample $\mathbf{X}$ with n observations. $\hat{\boldsymbol{\theta}}$ is a consistent estimator if its error, $E\{(\hat{\boldsymbol{\theta}}_n - \theta)^2\}$, tends to zero as $n \to \infty$.
- *Best estimator.* Let $\hat{\boldsymbol{\theta}} = g_n(\mathbf{X})$ be estimator of parameter θ. $\hat{\boldsymbol{\theta}}$ is best estimator if it minimizes the error $E\{(\hat{\boldsymbol{\theta}}_n - \theta)^2\}$.
- *Sufficient statistic.* A function $T(\mathbf{X})$ is a sufficient statistic for a parameter θ, if $T(\mathbf{X})$ contains all information about θ in the data set $\mathbf{X}$. That is, given the PDF $P\{\mathbf{x}_1 = x_1, \ldots, \mathbf{x}_n = x_n; \theta\}$, if $P\{\mathbf{x}_1 = x_1, \ldots, \mathbf{x}_n = x_n; \theta | T(\mathbf{X}\}$ *does not* depend on θ, then $T(\mathbf{X})$ is sufficient statistic for θ. (p322-323)
- *Efficient estimator.* An estimator is called efficient if it has the lowest possible variance. More specifically, we have first-order efficiency and second-order efficiency. The former means that an estimator is efficient w.r.t. the Cramer-Rao bound (see below), because the CR bound uses only first-order derivatives; the latter means that an estimator is efficient w.r.t. Bhattacharya bound, which is computed using the second-order derivative.

6.2 Parameter Estimation

Very important section with concepts directly applicable to machine learning. Specifically, it is shown why the Maximum Likelihood (ML) is so important, and from which mathematical framework it comes from. How can one obtain an unbiased estimator *with minimum variance*, or how one can measure the variance of a given unbiased estimator. It looks like, if one wants to use non-Gaussian RVs for machine learning, the tools presented in this section are fundamental to show how to find their parameters.

The concepts of Sufficient Statistic and Uniformly Minimum Variance Unbiased Estimator (UMVUE) are introduced. The latter is a desirable and most used estimator type, and the former leads to finding it. The most important theorems along the way are: *Factorization theorem* (helps finding sufficient statistic), the *Cramer-Rao Bound* (finds lower bound for variance of unbiased estimator), the *Bhattacharya Bound* (improves Cramer-Rao), and the *Rao-Blackwell theorem* (the finale: finds the UMVUE from sufficient statistic).

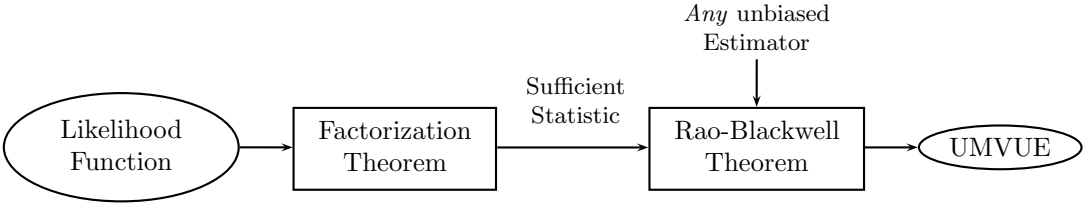

Figure 6.1: The relationship between the major concepts and theorems of this section (parameter estimation).

6.2.1 Notes

- **The Factorization Theorem** allows us to find suff. statistic (see also Examples 8-14, 8-15). Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a set of discrete RVs with PMF $P\{\mathbf{x}_1 = x_1, \ldots, \mathbf{x}_n = x_n; \theta\}$. Then, $T(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is sufficient for θ iff

$$P\{\mathbf{x}_1 = x_1, \ldots, \mathbf{x}_n = x_n; \theta | T(\mathbf{X})\} = h(x_1, \ldots, x_n) g_\theta(T\mathbf{x} = t) \tag{6.1}$$

  where $h(\cdot)$ is a nonnegative function of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ that does not depend on θ, and $g(\cdot)$ is a nonnegative function of θ and T only (and *not* of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in any other manner.)

- **Cramer-Rao (CR) Lower Bound** gives us the lower bound for the variance of an estimator $T(\mathbf{x})$. Let $T(\mathbf{x})$ be any unbiased estimator for θ under the joint PDF $f(\mathbf{x}_1 = x_1, \ldots, \mathbf{x}_n; \theta)$, denoted by $f(\mathbf{x}; \theta)$. Then,

$$\mathrm{Var}(T\{\mathbf{x}\}) \geq \frac{1}{E\left\{\left(\frac{\partial}{\partial\theta} \log f(\mathbf{x};\theta)\right)^2\right\}} = -\frac{1}{E\left\{\frac{\partial^2}{\partial\theta^2} \log f(\mathbf{x};\theta)\right\}} := \sigma^2_{\mathrm{CR}} \tag{6.2}$$

  provided that the following regularity conditions are satisfied

$$\frac{\partial}{\partial\theta} \int f(\mathbf{x};\theta) dx = \int \frac{\partial f(\mathbf{x};\theta)}{\partial\theta} dx = 0 \tag{6.3}$$

$$\frac{\partial}{\partial\theta} \int T(\mathbf{x}) f(\mathbf{x};\theta) dx = \int T(\mathbf{x}) \frac{\partial f(\mathbf{x};\theta)}{\partial\theta} dx. \tag{6.4}$$

  If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent, then

$$\sigma^2_{\mathrm{CR}} = \frac{1}{nE\left\{\left(\frac{\partial \log f(\mathbf{x}_i;\theta)}{\partial\theta}\right)^2\right\}} = \frac{1}{nE\left\{\frac{\partial^2 \log f(\mathbf{x}_i;\theta)}{\partial\theta^2}\right\}} \tag{6.5}$$

- **Cramer-Rao for Multiparameter Case**. Let $\underline{\theta} = (\theta_1, \ldots, \theta_m)$ be the parameter vector, and the corresponding estimator $\underline{T}(\mathbf{x}) = [T_1(\mathbf{x}), \ldots, T_m(\mathbf{x})]$. Then, it holds[1] $\mathrm{Cov}\{\underline{T}(\mathbf{x})\} \geq J^{-1}(\underline{\theta})$, where $J(\underline{\theta})$ represents the $m \times m$ Fisher matrix; i.e.,

$$J_{ij} = E\left\{\frac{\partial \log f(\mathbf{x};\underline{\theta})}{\partial\theta_i} \frac{\partial \log f(\mathbf{x};\underline{\theta})}{\partial\theta_j}\right\} \tag{6.6}$$

  The regularity conditions are similar—the $\partial\theta$ is replaced with $\partial\theta_i$.

- **First-order efficient**. An estimator whose variance is equal to σ^2_{CR} is called *first-order efficient* as the Cramer-Rao bound makes use of the first-order derivative.

- The pdf of an efficient estimator is from the *exponential family*. (p330)

- **Improvement on CR Bound: The Bhattacharya Bound**: The Bhattacharya Bound makes use of second order derivative. An estimator that attains this bound is called **second-order efficient**.

- **UMVUE**. The unbiased estimator for a parameter θ that has the lowest possible variance is called uniformly minimum variance unbiased estimator (UMVUE); the "the" in the former sentence is justified as the UMVUE is unique (p334). UMVUEs depend only on the sufficient statistic, and no other information about θ contained in the data set $\mathbf{X}$.

- The **Maximum-Likelihood Estimator (MLE)** $\hat{\boldsymbol{\theta}}_{ML}$ has very interesting properties, particularly for large n (p354).
  - If an efficient estimator exists, then *it is the MLE* (p330).
  - If MLE exists, then it is only a function of the sufficient statistic (see 8-283 and 8-284).
  - Assume that $\psi(\theta)$ is an unknown parameter depending on θ. Then, the MLE of ψ is given by $\psi(\hat{\boldsymbol{\theta}}_{ML})$.
  - If MLE is the unique solution to the likelihood equation, then under some additional restrictions and regularity conditions (see p354), we also have as $n \to \infty$: (i) $\hat{\boldsymbol{\theta}}_{ML}$ tends to a normal RV (ii) $\mathrm{Var}\hat{\boldsymbol{\theta}} \to \sigma^2_{\mathrm{CR}}$ (iii) $E(\hat{\boldsymbol{\theta}}_{ML}) \to \theta$; therefore $\frac{\hat{\boldsymbol{\theta}}_{ML} - \theta}{\sigma^2_{\mathrm{CR}}} \to N(0,1)$.

[1] $A \geq B$ in the sense that $A - B$ is a nonnegative-definite matrix

6.3 Confidence Interval Estimation

This section (p.307-317) is more like for evaluating the reliability of a particular parameter choice. E.g., what is the probability that the mean μ is within a certain interval? Or what is the interval (c_1, c_2) that contains μ with 95% confidence? Therefore I find it less applicable to signal processing and machine learning applications, but more appropriate like for testing hypotheses etc. In short: it's less interesting for me.

- General property: The minimum-length CI is such that $f(c_1) = f(c_2)$ (see p305 and Prob. 8-6), but finding this may not be easy (p314) therefore we may settle for a larger interval.
- **Mean**, η. The goal is estimating confidence interval (CI) for the mean η of $\mathbf{x}$ from *noisy* observations $x_1, ..., x_n$ drawn from RVs $\mathbf{x}_1, ..., \mathbf{x}_n$. The overall strategy is based on using the sample mean $\bar{\mathbf{x}}$ to find an *interval estimator* which is a pair of RVs which are functions of $\bar{\mathbf{x}}$, $(g_1(\bar{\mathbf{x}}), g_2(\bar{\mathbf{x}}))$, s.t. $P\{g_1(\bar{\mathbf{x}}) < \eta < g_2(\bar{\mathbf{x}})\} = \gamma$, where γ is *confidence coefficient*. We get the *interval estimation* by using the computed sample mean $\bar{x}$: $(g_1(\bar{x}), g_2(\bar{x}))$. This task is simplified when we assume the sample mean $\bar{\mathbf{x}}$ is Normal (which is true if $\mathbf{x}$ is Normal or approx true if n is large).
  - **Known Variance**, σ^2. $\bar{\mathbf{x}}$ is $\sim N(\eta, \sigma/\sqrt{n})$ in this case, therefore say that η is in the range $\bar{x} \pm z_{1-\delta/2}\sigma/\sqrt{n}$ with confidence γ, where $\delta = 1 - \gamma$ and z_u is the standard Normal distro's u-percentile.
  - **Unknown Variance**. We now use the (unbiased) sample variance estimator, $\mathbf{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^2$. We can be lazy and compute CI using the formula just above with s, *i.e.* $\bar{x} \pm z_{1-\delta/2}s/\sqrt{n}$ (which is OK for large n). Or, we can do things more properly. The RV $\frac{\bar{\mathbf{x}}-\eta}{\mathbf{s}/\sqrt{n}}$ has Student t distro with $n-1$ DoF. Denoting by t_u its percentiles, the CI that we are after is $\bar{x} \pm t_{1-\delta/2}\sigma/\sqrt{n}$.
  - **Nothing is known**. We use the elegant Tchebycheff Inequality, and our CI is $\bar{x} \pm \sigma/\sqrt{n\delta}$. (I assume that σ is sample variance? Not clear from book).
- **Zero-one event probability**, $P(A)$. A zero-one event (e.g. coin tossing or Republicans winning) is $\sim$ Bernoulli(p) and we aim to find CI for p. Since the mean of Bernoulli is p, we are effectively trying to estimate the mean of the distro, therefore we can use the stuff listed above. Specifically, for large n, our Bernoulli is approximated by $N(p, \sqrt{pq/n})$, therefore our CI is $\bar{x} \pm z_{1-\delta/2}\sqrt{pq}/\sqrt{n}$ with $p = \bar{x}$ and $q = 1 - p$.
- **Variance**, σ^2.
  - **Known mean**, η. The RV $\hat{\mathbf{v}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \eta)^2$ is an unbiased and consistent estimator of variance (p313), and $n\hat{\mathbf{v}}/\sigma^2$ is $\sim\sim \chi^2(n)$. We can then either build a CI by excluding equal tail prob. masses of $1 - \delta/2$, *i.e.* $\frac{n\hat{v}}{\chi^2_{1-\delta/2}(n)} < \sigma^2 < \frac{n\hat{v}}{\chi^2_{1-\delta/2}(n)}$. This interval *does not* have minimum length; but the determination of the latter is not simple (p313-314).
  - **Unknown mean**. Very similar to the case above (p314). We use σ^2's point estimator, $\mathbf{s}^2$, which is $\sim \chi^2(n-1)$. The CI is $\frac{(n-1)\hat{s}^2}{\chi^2_{1-\delta/2}(n-1)} < \sigma^2 < \frac{(n-1)\hat{s}^2}{\chi^2_{\delta/2}(n-1)}$.
- See p315-317 for estimating **Percentiles** (*i.e.* , x_u s.t. $F(x_u) = u$), **Distributions** ($\hat{F}(x)$) and interval estimates for a specific x value of Distributions.

6.4 Hypothesis Testing

Let H_0 be the null hypothesis that a parameter θ of a distribution equals θ_0, *i.e.* $H_0 : \theta = \theta_0$. The alternative hypothesis can be $H_1 : \theta \neq \theta_0$, or (more specifically) $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$. Hypothesis testing aims to establish whether there is sufficient evidence to reject H_0. The idea is to compute some quantity from the sample and test whether this quantity falls into the critical region, D_c.

Testing is based on the allowed errors; specifically, the probability *masses* $\alpha, \beta(\theta)$ (defs below)

	H_0 is really true	H_0 is really false
We reject H_0	Type I error (prob: $\alpha = P\{\mathbf{X} \in D_c \mid H_0\}$)	;))
We accept H_0	;)	Type II error (prob: $\beta(\theta) = P\{\mathbf{X} \notin D_c \mid H_1\}$)

Overall strategy to hypothesis testing is:

1. Select a test statistic $\mathbf{q} = g(\mathbf{X})$ (e.g.)
2. Find the critical region R_c where the density of $\mathbf{q}$ is negligible according to an α value.
3. Compute the value q using the observation vector $X = [x_1, x_2, \ldots, x_n]$
4. Reject H_0 if $q \in R_c$

Example. We want to test if the mean of $\mathbf{x}$ is η_0. The sample mean is $\bar{\mathbf{x}}$ and, under familiar assumptions, $\bar{\mathbf{x}} \sim N(\eta, \sigma/\sqrt{n})$

1. Select $\mathbf{q}$ as the standard (under H_0 assumption) NV: $\mathbf{q} = \frac{\bar{\mathbf{x}}-\eta_0}{\sigma/\sqrt{n}}$. Note that $\mathbf{q} \sim N(\eta_q, 1)$ where $\eta_q = \frac{\eta-\eta_0}{\sigma/\sqrt{n}}$
2. We set $\alpha = 0.05$, in which case the critical region for $R_c = \{q : z_{0.05} < q < z_{1-0.05}\}$
3. We compute $q(X)$ using the observation vector X
4. We reject $H_0 : \eta = \eta_0$ iff $q \in R_c$

The general formula above works across a variety of scenarios and tests. The example was for testing mean but the same strategy works for testing variance, (event) probability, distribution fit

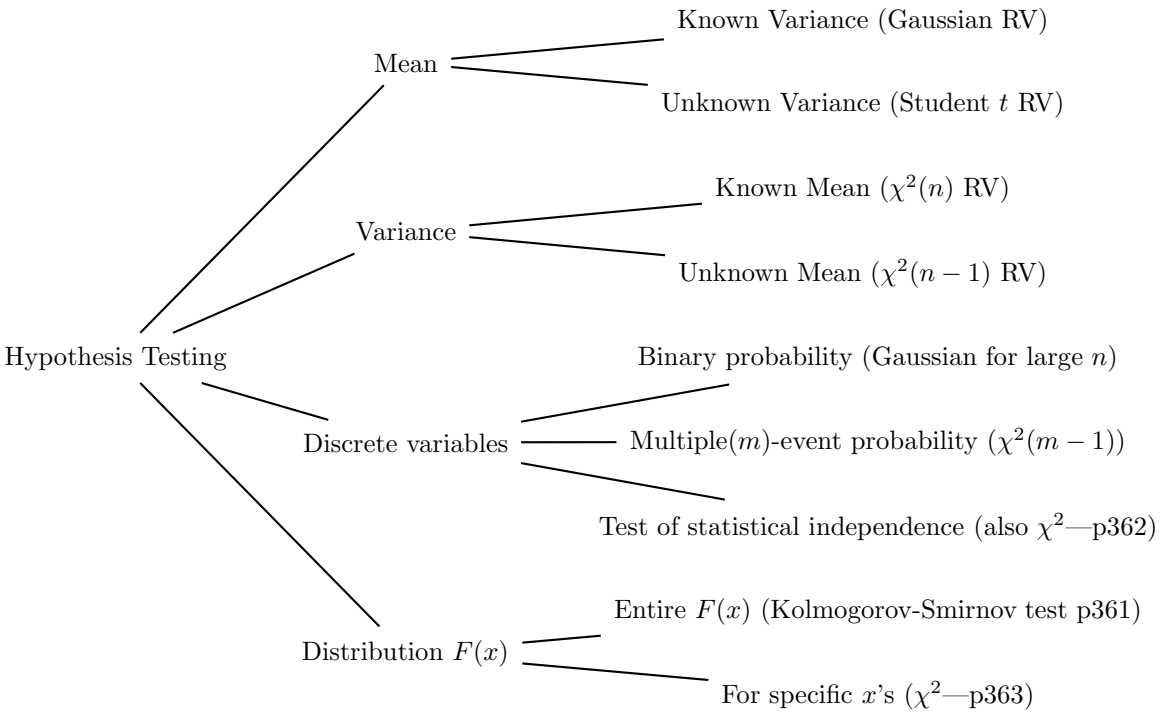

Figure 6.2: Summary (p358-364) of which tests are appropriate for the test of various quantities

Towards a motivation guide

6.5 Why transformations are useful

- By applying an orthonormal transformation (*a.k.a.* whitening) to a set of RVs we can easily compute the optimal
- How to optimally estimate: We do all our calculations in the space of the RVs; there we find an optimal (in some sense) estimators that are based on the hypothetical observation RVs (typically iid), and once we have the estimator, we apply our observations to this estimator (*i.e.* replace the RVs $\mathbf{x}_i$'s with the observed x_i's). Doing the estimation in the space of RVs gives us some theoretical guarantees under certain assumptions.

Towards a cheatsheet

6.6 Interesting RV Transformations

- If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are jointly normal, then $\mathbf{z} = \mathbf{x}_1 + \ldots \mathbf{x}_n$ is also normal (#257)
- If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent and each $\mathbf{x}_i$ is $N(0, 1)$, then $\mathbf{z} = \mathbf{x}_1 + \ldots \mathbf{x}_n$ is $\chi^2(n)$ (#259)
- If $\mathbf{x}$ is $\chi^2(n)$ and $\mathbf{y}$ is $\chi^2(m)$ and $\mathbf{x}, \mathbf{y}$ independent, then $\mathbf{z} = \mathbf{x} + \mathbf{y}$ is $\chi^2(n + m)$ (#260)

6.7 Useful Identities

- $\text{Var}(\mathbf{x}) = E\{\text{Var}(\mathbf{x}|\mathbf{y})\} + \text{Var}(E\{\mathbf{x}|\mathbf{y}\})$ (p337)

Sketches to solutions

6.8 Solutions to equations in the book

6.8.1 Error correction for CLT

It is really not obvious how the error correction for CLT is arrived at in #281. The general idea is to match the higher order moments of the approximation with that of the data. The approximation error between the (unknown) density $f(x)$ and the approximated normal density $f_n(x;0,\sigma)$ is:

$$\epsilon(x) = f(x) - f_n(x;0,\sigma). \tag{6.7}$$

The idea is to write the error (and thus $f(x)$) *i.t.o.* an orthogonal set of polynomials, viz. Hermite Polynomials $H_k(x)$ (see 7-126). Since they are orthogonal, they form an orthogonal set on the real line:

$$\int_{-\infty}^{\infty} e^{-x^2/2} H_n(x) H_m(x) dx = \begin{cases} n!\sqrt{2\pi} & n = m \\ 0 & n \neq m \end{cases} \tag{6.8}$$

Using those polynomials, one can approximate any function (including $\epsilon(x)$ or $f(x)$) *i.t.o.* an infinite series:

$$\epsilon(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} \sum_{k=3}^{\infty} C_k H_k\left(\frac{x}{\sigma}\right) \tag{6.9}$$

The sum starts from 3, as the moments of $\epsilon(x)$ up to order 2 are zero (I guess because the Gaussian approx. is sufficient to make those moments zero?) The book limits itself up to $4th$ order approx.

The idea is to find the coefficients C_3 and C_4 by matching the 3^{rd} and 4^{th} order moments of the unknown distro with that of the data (*** Be more pecise here). The third order moment of $\mathbf{x}$ is:

$$E_{f(x)}\{x^3\} = \int x^3 f(x) dx \tag{6.10}$$

The third order moment of the (approximated) normal density is zero (see 5-73), and therefore the error in terms of the third moments is:

$$E_{f(x)}\{x^3\} - E_{f_n(x)}\{x^3\} = \int x^3 f(x) dx - \int x^3 f_n(x;0,\sigma) dx = \int x^3 \left[f(x) - f_n(x;0,\sigma)\right] dx. \tag{6.11}$$

Since the content of the last brackets equals $\epsilon(x)$, and since $E_{f_n(x)}\{x^3\}$ is zero, the above equals $m_3 = E_{f(x)}\{x^3\}$ identity can be written as:

$$m_3 = \frac{1}{\sigma\sqrt{2\pi}} \int x^3 e^{-x^2/2\sigma^2} \sum_{k=3}^{\infty} C_k H_k\left(\frac{x}{\sigma}\right) dx \tag{6.12}$$

Now here is the tricky part that is not obvious in the book. Our goal is to find the coefficients C_k using the above equality. This is achieved by using the orthogonality of the Hermite polynomials. The third order Hermite Polynomial is $H_3(x) = x^3 - 3x$, and using (6.8) and doing a change of variables we see that:

$$\int e^{-x^2/2\sigma^2} \left(\frac{x^3}{\sigma^3} - 3\frac{x}{\sigma}\right)\left(\frac{x^3}{\sigma^3} - 3\frac{x}{\sigma}\right) dx = \tag{6.13}$$

$$= \int e^{-x^2/2\sigma^2} \frac{x^3}{\sigma^3}\left(\frac{x^3}{\sigma^3} - 3\frac{x}{\sigma}\right) dx - 3\int e^{-x^2/2\sigma^2} \frac{x}{\sigma}\left(\frac{x^3}{\sigma^3} - 3\frac{x}{\sigma}\right) dx \tag{6.14}$$

$$= 3!\sigma\sqrt{2\pi} \tag{6.15}$$

The second integral in (6.14) equals zero because $H_1(x) \propto x$ and $H_3(x) \perp H_1(x)$, and the first integral in (6.14) is proportional to m_3. More specifically, using (6.14) and (6.12), we arrive at the obscure derivation in the book:

$$m_3 = 3!\sigma^3 C_3. \tag{6.16}$$

Bibliography

[1] A. P. anb S. Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes.* McGraw - Hill, 2002.

[2] I. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products.* Academic Press, 1980.

[3] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions.* Dover, 1970.

Index