A Brief Review of Probability & Statistics

Draft of July 26, 2023

Thomas Li

thomli@upenn.edu

This document is a compilation of my notes from probability and statistics courses at the University of Pennsylvania. It will grow with time. Please contact me if you notice any errors.

Thanks to Professors Santosh Venkatesh, Hamed Hassani, Rajiv Gandhi, Karun Adusumilli, and Xu Cheng for their teaching and coverage of the material in this document.

CONTENTS

1 Discrete Probability **2**

1.1 The Probability Space 2
1.2 Properties of Discrete Spaces 3
1.3 Conditional Probability 4
1.4 Independence 5
1.5 Discrete Random Variables 6
1.6 Arithmetic Distributions 8
1.7 Probabilistic Bounds 10

2 Continuous Probability **12**

2.1 The One-Dimensional Continuous Probability Space 12
2.2 Univariate Distributions 13
2.3 The Multi-Dimensional Continuous Probability Space 15
2.4 Independence, Covariance, and Conditionality 16
2.5 Functions and Linear Transformations 18
2.6 The Normal Distribution 20
2.7 Other Distributions 22
2.8 Probabilistic Bounds 23

CHAPTER 1

DISCRETE PROBABILITY

§1.1 The Probability Space

The probability space abstracts a tangible chance experiment to a mathematical idealization. It is comprised of the probabilistic trinity $(\Omega, \mathcal{F}, \Pr)$, or sometimes just $(\Omega, \Pr)$. In words, the probability space can be defined as the sample space together with a probability function.

DEF. **Sample space:** the set of all outcomes of the chance experiment; denoted by Ω. An element (sample point) $\omega \in \Omega$ is an atomic outcome of the chance experiment.

DEF. **Event:** a measurable subset of the sample space; mathematically, $A \subseteq \Omega$. Outside the abstraction, an event is a collection of outcomes.

DEF. **Family of events:** the set of events of interest; denoted by $\mathcal{F}$.

DEF. **Probability measure (function):** a set function that assigns to every event a real value satisfying the probability axioms; mathematically, $\Pr : \mathcal{F} \to [0, 1], A \mapsto \Pr(A)$.

DEF. **Probability axioms:**

- **Positivity:** $\Pr(A) \geq 0, \forall A \in \mathcal{F}$. All events have positive probability.
- **Normalization:** $\Pr(\Omega) = 1$. The probability assigned to the entire sample space is 1.
- **Additivity:** For disjoint events $A_1, A_2, \ldots \in \mathcal{F}$,

$$\Pr(A_1 \cup A_2 \cup \ldots) = \Pr(A_1) + \Pr(A_2) + \ldots$$

§1.2 Properties of Discrete Spaces

DEF. **Discrete space:** a sample space that is countable (finite or denumerably infinite). Mathematically, $\Omega = \{\omega_i \mid i \geq 1\}$.

DEF. **Generic probability functions:** for a discrete space Ω,

- $\Pr : \{\omega_i\} \mapsto p(i) \geq 0, \forall i \geq 1$ (positivity); call $p(\cdot)$ the **mass function**,
- $\Pr(\Omega) = \sum_i \Pr\{\omega_i\} = \sum_i p(i) = 1$ (additivity and normalization),
- For an event $A = \{\omega_{i_1}, \omega_{i_2}, \ldots\}$, $Pr : A \mapsto p(i_1) + p(i_2) + \cdots$ (additivity).

DEF. **Uniform spaces:** a finite sample space Ω such that each outcome ω is equally likely. For a generic event A, the probability function is combinatorial:

$$\Pr(\omega) = \frac{1}{|\Omega|} \text{ and } \Pr(A) = \frac{|A|}{|\Omega|}$$

.

We now list a few theorems that follow from the probability axioms:

- **Empty Set**: $\Pr(\emptyset) = 0$
- **Complements**: $\Pr(A^\complement) = 1 - \Pr(A)$
- **Monotonicity**: $A, B \in \mathcal{F}$ and $A \subseteq B \implies \Pr(A) \leq \Pr(B)$
- **Boundedness**: $A \in \mathcal{F} \implies 0 \leq \Pr(A) \leq 1$
- **DeMorgan's**: $\left(\bigcup_i A_i\right)^\complement = \bigcap_i A_i^\complement$ and $\left(\bigcap_i A_i\right)^\complement = \bigcup_i A_i^\complement$
- **Boole's inequality (union bound)**: $\Pr\left(\bigcup_i A_i\right) \leq \sum_i \Pr(A_i)$
- **Boole's sieve**: $\sum_i \Pr(A_i) < 1 \implies \bigcap_i A_i^\complement \neq \emptyset$

THM. **Principle of Inclusion-Exclusion:**

- *Simplified.* For two events A, B,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

- *Generalized.* For n events $A_1, \ldots, A_n$ and $I_j = \Pr(A_1 \cap \cdots \cap A_j) + \cdots + \Pr(A_{n-j+1} \cap \cdots \cap A_n)$ (the sum of all j-wise intersections),

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = I_1 - I_2 + I_3 - \cdots + (-1)^{n-1} I_n$$

§1.3 Conditional Probability

Def. **Conditional probability:** for two events A, B where $\Pr(B) \neq 0$,

$$\Pr(A \mid B) \coloneqq \frac{\Pr(A \cap B)}{\Pr(B)}$$

Some corollaries that follow from this definition:

- $\Pr(B \mid B) = 1$
- $\Pr(A^{\complement} \mid B) = 1 - \Pr(A \mid B)$
- $\Pr(A \cap B) = \Pr(A \mid B)\Pr(B) = \Pr(B \mid A)\Pr(A)$
- $A \perp B \iff \Pr(A \mid B) = \Pr(A) \iff \Pr(B \mid A) = \Pr(B)$

Thm. **Chain rule:**

- *Simplified.* For three events A, B, C,

$$\Pr(A \cap B \cap C) = \Pr(A \mid B \cap C) \times \Pr(B \mid C) \times \Pr(C)$$

- *Generalized.* For n events $A_1, \ldots, A_n$,

$$\begin{aligned}\Pr(A_1 \cap \cdots \cap A_n) = &\Pr(A_1 \mid A_2 \cap \cdots \cap A_n) \times \Pr(A_2 \mid A_3 \cap \cdots \cap A_n) \\ &\times \cdots \times \Pr(A_{n-1} \mid A_n) \times \Pr(A_n)\end{aligned}$$

Thm. **Bayes' Theorem:** for two events A, B,

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A)}$$

Thm. **Law of total probability:**

- *Simplified.* For two events E, A,

$$\begin{aligned}\Pr(E) &= \Pr(E \mid A)\Pr(A) + \Pr(E \mid A^{\complement})\Pr(A^{\complement}) \\ &= \Pr(E \cap A) + \Pr(E \cap A^{\complement})\end{aligned}$$

- *Generalized.* For events $E, A_1, \ldots, A_n$ where $A_1, \ldots, A_n$ partition Ω,

$$\begin{aligned}\Pr(E) &= \Pr(E \mid A_1)\Pr(A_1) + \cdots + \Pr(E \mid A_n)\Pr(A_n) \\ &= \Pr(E \cap A_1) + \cdots + \Pr(E \cap A_n)\end{aligned}$$

§1.4 Independence

DEF. **Independence:** two events A, B are independent (write $A \perp B$) iff

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

Some corollaries that follow from this definition:

- $A \perp B \iff \Pr(A \mid B) = \Pr(A) \iff \Pr(B \mid A) = \Pr(B)$
- $A \perp B \iff A \perp B^{\complement} \iff A^{\complement} \perp B^{\complement} \iff A^{\complement} \perp B$
- $A \perp B \implies \Pr(A \cup B) = 1 - (1 - \Pr(A))(1 - \Pr(B))$

DEF. **Pairwise independence:** events $A_1, A_2, \ldots, A_n$ are *pairwise* independent iff for all $i \neq j$, $A_i \perp A_j$.

DEF. **Mutual Independence:** events $A_1, A_2, \ldots, A_n$ are *mutually* independent iff for all $I \subseteq \{1, 2, \ldots, n\}$,

$$\Pr\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \Pr(A_i).$$

Note pairwise independence $\not\implies$ mutual independence, but mutual independence $\implies$ pairwise independence.

Sequence of indepedent trials: Consider a chance experiment of n independent trials, where each trial i has the set of outcome T_i and a marginal mass function $p_i(k)$. Then, the sample space Ω can be described as

$$\Omega = T_1 \times T_2 \times \cdots \times T_n$$

and an outcome of Ω can be described as

$$\omega = (t_1, \ldots, t_n),$$

where $t_i \in T_i$ is the outcome of the ith trial. Due to independence, the probability function of the space is

$$\Pr(\omega) = \Pr((t_1, \ldots, t_n)) = p_1(t_1) \times \cdots \times p_n(t_n)$$

DEF. **Conditional independence:** events A, B are conditionally independent given C iff

$$\Pr(A \cap B \mid C) = \Pr(A \mid C) \Pr(B \mid C)$$

Note that independence $\not\implies$ conditional independence.

§1.5 Discrete Random Variables

Def. **Random variable:** a quantitative measure of a chance experiment. May think of an r.v. as a function $X : \Omega \to \mathbb{R}, \omega \mapsto X(\omega) \in \mathbb{R}$.

- $\{X = k\} \equiv \{\omega \mid X(\omega) = k\} \subseteq \Omega$, the set of all outcomes ω that map to the realization k.

Def. **Arithmetic random variable:** random variables that only take on integer values: $X : \Omega \to \mathbb{Z}$.

Def. **Support:** $\text{Supp}(X) \equiv \{X(\omega) : \omega \in \Omega\}$, the set of all possible realizations of an r.v. X.

Def. **Probability distribution:** Let $p : \text{Supp}(X) \to [0, 1], k \mapsto \Pr(\{X = k\}) = p(k)$ be the mass function of the r.v. X. The probability distribution of X is $p(k), \forall k \in \text{Supp}(X)$.

- $X \sim p(k)$ is notation for a r.v. X distributed according to probability function $p(k)$.

Def. **Expectation:** for an arithmetic r.v. X,

$$\mathbb{E}[X] = \mu_X \coloneqq \sum_{k \in X} k \cdot p(k), \quad \text{if convergent.}$$

Some properties of expectation:

- $\mathbb{E}[X]$ is a constant.
- Thm. **Linearity of expectation:** for constants a, b, c and r.v.s X, Y,

  $$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c.$$

  It follows that for any r.v.s $X_1, \ldots, X_n$,

  $$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

- Thm. For a non-negative r.v. $X \sim p(k)$ (that is, $\text{Supp}(X) \subseteq \mathbb{N}$),

  $$\mathbb{E}[X] = \sum_{k=1}^{\infty} k \cdot p(k) = \sum_{k=1}^{\infty} \Pr(X \geq k)$$

Def. **Moment:** the rth moment of an r.v. X is given by $\mathbb{E}[X^r]$.

Def. **Variance:** for an r.v. X,

$$\text{Var}(X) = \sigma_X^2 \coloneqq \mathbb{E}[(X - \mu_X)^2] = \sum_{k \in \text{Supp}(X)} (k - \mu_X)^2 \cdot p(k), \quad \text{if convergent.}$$

Variance measures the spread of a probability distribution.

Some properties of variance:

- $\mathrm{Var}(X)$ is a constant.
- Equivalent formula: $\sigma_X^2 = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}[X^2] - \mu_X^2$.
- $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$, for constants a, b and r.v. X
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}$
- Thm. **Variance of sum of independent r.v.s:**

$$X \perp Y \implies \mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

  Generalized, if $X_1, X_2, \ldots, X_n$ are pairwise independent r.v.s,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

Def. **Standard deviation:** $\mathrm{sd}(X) = \sigma_X = \sqrt{\sigma_X^2}$.

Some properties of standard deviation:

- Has the same units as X
- $\mathrm{sd}(aX + b) = |a|\,\mathrm{sd}(X)$, for constants a, b and r.v. X

Def. **Independence of random variables:** Two r.v.s X, Y are independent (write $X \perp Y$) iff for all $x \in \mathrm{Supp}(X), y \in \mathrm{Supp}(Y)$,

$$\Pr(X = x \cap Y = y) = \Pr(X = x) \cdot \Pr(Y = y)$$

Def. **Mutual independence of random variables:** A collection of n r.v.s $X_1, \ldots, X_n$ are mutually independent iff for all $I \subseteq \{1, \ldots, n\}$,

$$\Pr\left[\bigcap_{i \in I} X_i = x_i\right] = \prod_{i \in I} \Pr\left(X_i = x_i\right)$$

Def. **Conditional expectation:** for two r.v.s X, Y

$$\mathbb{E}(X \mid Y = y) = \sum_{x \in \mathrm{Supp}(X)} x \cdot \Pr(X = x \mid Y = y)$$

Thm. **Law of total expectation:** for two r.v.s X, Y

$$\mathbb{E}[X] = \sum_{y \in \mathrm{Supp}(Y)} \mathbb{E}[X \mid Y = y] \cdot p_Y(y)$$

§1.6 Arithmetic Distributions

Def. **Bernoulli:** A Bernoulli r.v. can be likened to a coin flip (either a success or failure) with success probability p. An r.v. $X \sim Bernoulli(p)$ for $p \in [0,1]$ has the following properties:

- $\text{Supp}(X) = \{0, 1\}$
- $\Pr(X = 1) = p$ and $\Pr(X = 0) = 1 - p$
- $\mathbb{E}[X] = p$
- $\text{Var}(X) = p(1 - p)$

Def. **Indicator function:** a Bernoulli r.v. defined upon an event A:

$$\mathbb{1}_A(\omega) = \begin{cases} 1, \text{ if } \omega \in A \\ 0, \text{ o.w.} \end{cases}$$

Def. **Binomial:** A binomial r.v. can be thought of as the number of successes among n i.i.d. $Bernoulli(p)$ trials. An r.v. $X \sim Binomial(n, p)$ for $n \in \mathbb{Z}^+, p \in [0,1]$ has the following properties:

- $\text{Supp}(X) = \{k \in \mathbb{Z} \mid 0 \leq k \leq n\}$
- $\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$
- $\mathbb{E}[X] = np$
- $\text{Var}(X) = np(1 - p)$

Def. **Poisson:** A Poisson r.v. provides a good approximation of a $Binomial(n, p)$ r.v. when $p \ll 1, n \gg 1$ (rare events), in which the parameter $\lambda = np$. An r.v. $X \sim Poisson(\lambda), \lambda > 0$ has the following properties:

- $\text{Supp}(X) = \mathbb{N}$
- $\Pr(X = k) = Po(k; \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}$
- $\mathbb{E}[X] = \lambda$
- $\text{Var}(X) = \lambda$
- Characteristic property: $kPo(k; \lambda) = \lambda Po(k - 1; \lambda)$
- Thm. **Sum of Poissons:** Consider a set of n Poisson r.v.s $\{X_i \sim Poisson(\lambda_i), 1 \leq i \leq n\}$. Let $S_n = X_1 + \cdots + X_n$. Then, $S_n \sim Poisson(\lambda_1 + \cdots + \lambda_n)$.

Proof. Let's start with two Poisson r.v.s X_1, X_2 with respective parameters λ_1, λ_2 and mass functions $p_1(x_1), p_2(x_2)$. To derive the mass function $q(k)$ of $S = X_1 + X_2$, we are concerned with the event $\{S = k\} = \{X_1 + X_2 = k\} = \{(x_1, x_2) \mid x_1 + x_2 = k\}$. By additivity, we have the convolution[a]:

$$\begin{aligned} q(k) &= \sum_{x=0}^{k} p_1(x)p_2(k-x) \\ &= \sum_{x=0}^{k} e^{-\lambda_1}\frac{\lambda_1^x}{x!} \cdot e^{-\lambda_2}\frac{\lambda_2^{k-x}}{(k-x)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{x=0}^{k}\frac{k!}{x!(k-x)!}\lambda_1^x\lambda_2^{k-x} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{x=0}^{k}\binom{k}{x}\lambda_1^x\lambda_2^{k-x} \\ &= e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^k}{k!} \end{aligned}$$

This is precisely the mass function of a Poisson r.v. with parameter $\lambda_1 + \lambda_2$. The proof extends to the sum of n Poisson r.v.s by induction: $S_n = S_{n-1} + X_n$. ■

[a] Given two functions $\alpha(\cdot), \beta(\cdot)$, a convolution is defined to be $\alpha * \beta(k) \coloneqq \sum_j \alpha(j)\beta(k-j)$.

Def. **Geometric**[1]**:** A geometric r.v. is a waiting time distribution: specifically, the number of $Bernoulli(p)$ failures before the first success. An r.v. $X \sim Geometric(p), p \in [0, 1]$ has the following properties:

- $\text{Supp}(X) = \mathbb{N}$
- $\Pr(X = k) = p \cdot (1-p)^k$
- $\mathbb{E}[X] = \frac{1-p}{p}$
- $\text{Var}(X) = \frac{1-p}{p^2}$
- **Memoryless property:** $\Pr(X = n + k \mid X > k) = \Pr(X = n)$

Def. **Negative Binomial:** The negative binomial generalizes the geometric distribution: it is the number of $Bernoulli(p)$ failures before the nth success. An r.v. $X \sim NB(n, p)$ has the following properties:

- $\text{Supp}(X) = \mathbb{N}$

[1] Some textbooks define a geometric r.v. as the number of trials—rather than just failures—needed for the first success. Then, $\text{Supp}(X) = \mathbb{Z}^+$, $\Pr(k) = p \cdot (1-p)^{k-1}$, $\mathbb{E}[X] = \frac{1}{p}$, $\text{Var}(X) = \frac{1-p}{p^2}$.

- $\Pr(X = k) = w_n(k; p) = \binom{n+k-1}{k}(1-p)^k p^n$
- $\mathbb{E}[X] = n \cdot \frac{1-p}{p}$
- $\mathrm{Var}(X) = n \cdot \frac{1-p}{p^2}$

§1.7 Probabilistic Bounds

DEF. **Markov's inequality:** Markov provides an upper bound on the probability that a non-negative r.v. X is greater than some constant $\tau > 0$:

$$\Pr(X \geq \tau) \leq \frac{\mathbb{E}[X]}{\tau}$$

Proof. Markov follows from the definition of expectation:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_k kp(k) \\
&= \sum_{k<\tau} kp(k) + \sum_{k\geq\tau} kp(k) \\
&\geq \sum_{k\geq\tau} kp(k) \\
&\geq \tau \sum_{k\geq\tau} p(k) \\
&= \tau \Pr(X \geq \tau).
\end{aligned}$$

Rearranging gives us $\Pr(X \geq \tau) \leq \frac{\mathbb{E}[X]}{\tau}$. ■

DEF. **Chebyshev's inequality:** Chebyshev provides an upper bound on the probability that any r.v. X falls outside a range $\tau, \epsilon > 0$ from its mean. Three equivalent formulations:

$$\Pr\left(|X - \mu_X| \geq \tau\right) \leq \frac{\sigma_X^2}{\tau^2}$$

$$\Pr\left(\left|\frac{S_n}{n} - \mu_X\right| \geq \epsilon\right) \leq \frac{\sigma_X^2}{n\epsilon^2}$$

$$\Pr\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \leq \frac{1}{4n\epsilon^2}$$

where $S_n = X_1 + \cdots + X_n$ for $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Bernoulli(p)$, $\mu_X = \mathbb{E}[X_i]$, and $\sigma_X^2 = \mathrm{Var}[X_i]$. The third equation follows from $\max_{p\in[0,1]} p(1-p) = 1/4$.

Proof. Chebyshev follows from Markov:

$$\begin{aligned}\Pr\left(|X-\mu_X| \geq \tau\right) &= \Pr\left((X-\mu_X)^2 \geq \tau^2\right) \\ &\leq \frac{\mathbb{E}\left[(X-\mu_X)^2\right]}{\tau^2} \\ &= \frac{\mathrm{Var}(X)}{\tau^2}\end{aligned}$$

■

Applying Chebyshev to sampling. Suppose we wanted to estimate the fraction p of a population that has some (dichotomous) property with an error tolerance of ϵ and a confidence of $1-\delta$. If we sample n people, each one modeled as a $Bernoulli(p)$ r.v., our estimate of p is then $\frac{S_n}{n}$, where $S_n = X_1 + \cdots + X_n$, the number of people sampled with the desired property. Our error and confidence constraints are modeled as such:

$$\Pr\left(\left|\frac{S_n}{n} - p\right| < \epsilon\right) \geq 1-\delta,$$

or equivalently (by complement),

$$\Pr\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \leq \delta.$$

Applying Chebyshev's (the third formulation listed above), we have that

$$\Pr\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \leq \frac{1}{4n\epsilon^2} \leq \delta$$

and so we choose $n \geq \frac{1}{4\delta\epsilon^2}$. Note that as $n \to \infty$, $\frac{S_n}{n} \xrightarrow{\Pr} p$ (the weak law of large numbers).

CHAPTER 2

CONTINUOUS PROBABILITY

§2.1 The One-Dimensional Continuous Probability Space

We extend the following discrete concepts to the continuous space:

- **Sample space**: $\Omega = \mathbb{R}$
- **Events** are intervals (a, b). Infinitesimal event: $(x, x + \mathrm{d}x)$; generic event: $\mathbb{A} \subseteq \mathbb{R}$
- **Probability measure**, where $f(x)$ is density function (think of $\mathrm{d}x$ as a length such that $f(x)\,\mathrm{d}x$ is a probability mass):
  - $\Pr : (x, x + \mathrm{d}x) \mapsto f(x)\,\mathrm{d}x$
  - $\Pr : (a, b) \mapsto \int_a^b f(x)\,\mathrm{d}x$
  - $\Pr : \mathbb{A} \mapsto \int_{\mathbb{A}} f(x)\,\mathrm{d}x$
- **Random variable** $X : \Omega \to \mathbb{R}$. $X \sim f(x)$ denotes that the r.v. X is distributed according to the probability density function $f(x)$.
- **Expectation** $\mathbb{E}[X] = \mu_X \coloneqq \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$, if convergent
- **Variance** $\mathrm{Var}(X) = \sigma_X^2 \coloneqq \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)\,\mathrm{d}x$, if convergent

Def. **Probability density function (pdf):** the density function f is a non-negative (positivity) integrable function with unit area (normalization).

Def. **Cumulative distribution function (cdf):** the distribution function F associated with the density function f is defined as such

$$F(x) \coloneqq \Pr : (-\infty, x] \mapsto \int_{-\infty}^{x} f(u)\,\mathrm{d}u$$

Note that a distribution function is monotonically increasing and is continuous.

By the Fundamental Theorem of Calculus,

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) = \frac{\mathrm{d}}{\mathrm{d}x}\int_{-\infty}^{x} f(u)\,\mathrm{d}u = f(x)$$

To measure an event associated with r.v. $X \sim f(x), F(x)$:

- $\Pr(a \leq X \leq b) = \Pr((a,b)) = \int_a^b f(x)\,\mathrm{d}x = \int_a^b \mathrm{d}F(x) = F(b) - F(a)$
- $\Pr(X \in \mathbb{A}) = \Pr(\mathbb{A}) = \int_{\mathbb{A}} f(x)\,\mathrm{d}x = \int_{\mathbb{A}} \mathrm{d}F(x)$

§2.2 Univariate Distributions

Def. **Unit uniform:** an r.v. $X \sim Uniform(0,1)$ has the following properties:

- $\mathrm{Supp}(X) = (0,1)$
- pdf $f(x)$ and cdf $F(x)$:

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{o.w.} \end{cases}, \qquad F(x) = \begin{cases} 0, & x \leq 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x \geq 1 \end{cases}$$

- $\mathbb{E}[X] = \frac{1}{2}$
- $\mathrm{Var}(X) = \frac{1}{12}$

Def. **Uniform:** within the support (a,b), all intervals of the same length are equally probable. An r.v. $X \sim Uniform(a,b)$ has the following properties:

- $\mathrm{Supp}(X) = (a,b)$, for real numbers $a < b$
- pdf $f(x)$ and cdf $F(x)$:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{o.w.} \end{cases}, \qquad F(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x \geq b \end{cases}$$

- $\mathbb{E}[X] = \frac{a+b}{2}$
- $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$
- Transformation from unit uniform: $X = (b-a)Y + a$, where $Y \sim Uniform(0,1)$.

DEF. **Unit exponential:** an r.v. $X \sim Exponential(1)$ has the following properties:

- $\text{Supp}(X) = \mathbb{R}^+$
- pdf $f(x) = e^{-x}, x > 0$
- cdf $F(x) = 1 - e^{-x}, x > 0$:
- $\mathbb{E}[X] = 1$
- $\text{Var}(X) = 1$

DEF. **Exponential:** an r.v. $X \sim Exponential(\alpha)$ has the following properties:

- $\text{Supp}(X) = \mathbb{R}^+$
- pdf $f(x) = \alpha e^{-\alpha x}, x > 0$
- cdf $F(x) = 1 - e^{-\alpha x}, x > 0$
- $\Pr(X > c) = e^{-\alpha c}$, for some constant c
- $\mathbb{E}[X] = \frac{1}{\alpha}$
- $\text{Var}(X) = \frac{1}{\alpha^2}$
- **Memoryless property**: $\Pr(X > s + t \mid X > s) = \Pr(X > t)$
- Transformation from unit exponential: $X = \frac{1}{\alpha}Y$, where $Y \sim Exponential(1)$

DEF. **Standard normal:** an r.v. $X \sim \mathcal{N}(0, 1)$ has the following properties:

- $\text{Supp}(X) = \mathbb{R}$
- pdf

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

- cdf $\Phi(x) = \int_{-\infty}^{x} \phi(u)\, \mathrm{d}u$
- $\mathbb{E}[X] = 0$
- $\text{Var}(X) = 1$

DEF. **Normal:** an r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$ has the following properties:

- $\text{Supp}(X) = \mathbb{R}$
- pdf

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

- cdf $F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u$
- $\mathbb{E}[X] = \mu$
- $\mathrm{Var}(X) = \sigma^2$
- $\Pr(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$, where Φ is the standard normal cdf.

Proof. We standardize the normal r.v.:

$$\Pr(X \leq x) = \Pr\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Pr\left(Y \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

■

- Transformation from standard normal: $X = \sigma Y + \mu$, where $Y \sim \mathcal{N}(0,1)$

§2.3 The Multi-Dimensional Continuous Probability Space

We extend the following one-dimensional concepts to the two-dimensional space:

- **Sample space:** $\Omega = \mathbb{R}^2$
- **Events** are areas $(a_1, b_1) \times (a_2, b_2)$. Infinitesimal event: $(x, x+\mathrm{d}x) \times (y, y+\mathrm{d}y)$; generic event: $\mathbb{A} \subseteq \mathbb{R}^2$.
- **Probability measure**, where $f(x,y)$ is the **joint density function**:
  - $\Pr : (x, x+\mathrm{d}x) \times (y, y+\mathrm{d}y) \mapsto f(x,y)\,\mathrm{d}x\,\mathrm{d}y$
  - $\Pr : \mathbb{A} \mapsto \iint_{\mathbb{A}} f(x,y)\,\mathrm{d}y\,\mathrm{d}x$
- **Joint distribution function:**

$$F(x,y) \coloneqq \Pr : (-\infty, x] \times (\infty, y] \mapsto \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,\mathrm{d}v\,\mathrm{d}u = \Pr(X \leq x, Y \leq y)$$

$$f(x,y) = \frac{\partial^2}{\partial x\,\partial y} F(x,y)$$

Def. **Marginal density:** describes the distribution governing an individual variable from a joint distribution. Given $(X,Y) \sim f(x,y), F(x,y)$, $X \sim f_X(x), F_X(x)$ (symmetric for Y), where

$$f_X(x) \coloneqq \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \quad \text{and} \quad F_X(x) \coloneqq \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u$$

To ground the idea of marginals, it can be helpful to consider the following: $\Pr(X \in \mathbb{A}) = \Pr(X \in \mathbb{A} \cap Y \in \mathbb{R}) = \int_{x\in\mathbb{A}}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int_{\mathbb{A}} f_x(x)\,\mathrm{d}x$

From the marginal distribution, we also derive the expectation and variance of an individual variable in a two-dimensional setting:

$$\mathbb{E}[X] = \mu_X \coloneqq \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\,\mathrm{d}y\,\mathrm{d}x$$
$$\mathrm{Var}(X) = \sigma_X^2 \coloneqq \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)^2 f(x,y)\,\mathrm{d}y\,\mathrm{d}x$$

Now to the nth-dimensional space:

- **Sample space:** $\Omega = \mathbb{R}^n$
- **Events:** $\mathbb{A} \subseteq \mathbb{R}^n$
- **Probability measure**, where $f(x_1,\ldots,x_n)$ is the joint density function and $F(x_1,\ldots,x_n)$ is the joint distribution function:

$$\Pr : \mathbb{A} \mapsto \int\cdots\int_{\mathbb{A}} f(x_1,\ldots,x_n)\,\mathrm{d}x_n\cdots\mathrm{d}x_1 = \int\cdots\int_{\mathbb{A}} \mathrm{d}F(x_1,\ldots,x_n)$$
$$\Pr : \mathbb{A} \mapsto \int\cdots\int_{\mathbb{A}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int\cdots\int_{\mathbb{A}} \mathrm{d}F(\boldsymbol{x}) \qquad \text{(in vector notation)}$$

- **Marginal density:** Given $\boldsymbol{X} \sim f(\boldsymbol{x}), F(\boldsymbol{x})$,

$$X_1 \sim f_1(x_1) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\boldsymbol{x})\,\mathrm{d}x_n\cdots\mathrm{d}x_2$$
$$\mathbb{E}[X_1] = \mu_1 \coloneqq \int_{-\infty}^{\infty} x_1 f_1(x_1)\,\mathrm{d}x_1$$
$$\mathrm{Var}(X_1) = \sigma_1^2 \coloneqq \int_{-\infty}^{\infty} (x_1-\mu_1)^2 f_1(x_1)\,\mathrm{d}x_1$$

§2.4 Independence, Covariance, and Conditionality

Def. **Independence:** Random variables $X, Y \sim f(x,y)$ are independent (write $X \perp Y$) are independent iff $f(x,y) = f_X(x)\cdot f_Y(y)$ (equivalently, $F(x,y) = F_X(x)\cdot F_Y(y)$).

Events determined solely by X are independent of events determined solely by Y if X, Y are independent:

$$\Pr(X\in\mathbb{A}, Y\in\mathbb{B}) = \Pr((x,y)\in\mathbb{A}\times\mathbb{B}) = \Pr(x\in\mathbb{A})\cdot\Pr(y\in\mathbb{B}) = \int_{\mathbb{A}} f_X(x)\,\mathrm{d}x\cdot\int_{\mathbb{B}} f_Y(y)\,\mathrm{d}y$$

Def. **Covariance:** For r.v.s $X, Y \sim f(x,y)$,

$$\begin{aligned}\mathrm{Cov}(X,Y) &\coloneqq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y)\,\mathrm{d}y\,\mathrm{d}x\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f(x,y)\,\mathrm{d}y\,\mathrm{d}x - \mu_X\mu_Y\\ &= \mathbb{E}[XY] - \mu_X\mu_Y\end{aligned}$$

Def. **Covariance matrix:**

$$\mathrm{Cov}(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\intercal] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}^\intercal]$$

If $X \perp Y$, then $\mathrm{Cov}(X, Y) = 0$. But this relationship does not hold in the other direction.

Def. **Conditional density:** for $X, Y \sim f(x, y)$, the conditional density of X given $Y = y$ is

$$f_X(x \mid Y = y) \coloneqq \frac{f(x, y)}{f_Y(y)}$$

Derivation from first principles:

$$\begin{aligned}\Pr(x < X < x + \mathrm{d}x \mid y < Y < y + \mathrm{d}y) &= \frac{\Pr(x < X < x + \mathrm{d}x, y < Y < y + \mathrm{d}y)}{\Pr(y < Y < y + \mathrm{d}y)} \\ &= \frac{f(x, y)\,\mathrm{d}x\,\mathrm{d}y}{f_Y(y)\,\mathrm{d}y} \\ &\coloneqq f_X(x \mid Y = y)\,\mathrm{d}x\end{aligned}$$

Given that $Y = y$, the probability that $X \in \mathbb{A}$ is

$$\Pr(X \in \mathbb{A} \mid Y = y) = \int_{\mathbb{A}} f_X(x \mid Y = y)\,\mathrm{d}x$$

We may also condition on an event:

$$\Pr(X \in \mathbb{A} \mid X \in \mathbb{B}) = \frac{\int_{\mathbb{A}\cap\mathbb{B}} f_X(x)\,\mathrm{d}x}{\Pr(X \in \mathbb{B})} = \int_{\mathbb{A}} f_{X|X\in\mathbb{B}}(x)\,\mathrm{d}x$$

where

$$f_{X|X\in\mathbb{B}}(x) = \frac{f_X(x)}{\Pr(X \in \mathbb{B})}, \text{ if } X \in \mathbb{B} \quad (0 \text{ o.w. })$$

Thm. **Law of Total Probability:** for a partition $A_1, \ldots, A_n$,

$$f_X(x) = \sum_{i=1}^{n} \Pr(A_i) f_X(x \mid A_i)$$

Def. **Conditional expectation:**

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid Y = y] \coloneqq \int_{-\infty}^{\infty} x f_X(x \mid y)\,\mathrm{d}x$$

$$\mathbb{E}[X \mid \mathbb{A}] \coloneqq \int_{-\infty}^{\infty} x f_X(x \mid \mathbb{A})\,\mathrm{d}x$$

THM. **Law of Iterated Expectation:** for a partition $A_1, \ldots, A_n$,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X \mid A_i] \Pr[A_i]$$

We may also condition on another random variable Y:

$$\mathbb{E}[X] = \sum_{y \in \text{Supp}(Y)} \mathbb{E}[X \mid Y = y] f_Y(y) = \mathbb{E}[\mathbb{E}[X \mid Y]]$$

§2.5 Functions and Linear Transformations

DEF. **Diffeomorphism:** a function that is differentiable and invertible.

Diffeomorphic transformation (1D). Consider an r.v. $X \sim f_X(x), F_X(x)$ and another r.v. $Y = \psi(X) \sim f_Y(y), F_Y(y)$ defined in terms of X by the diffeomorphism $\psi : \mathbb{R} \to \mathbb{R}$. Then, the density of Y is given by

$$f_Y(y) = f_X(\psi^{-1}(y)) J_{\psi^{-1}}(y),$$

where $J_{\psi^{-1}}(y)$ is the Jacobian $\left| \frac{\mathrm{d}}{\mathrm{d}y} \psi^{-1}(y) \right|$.

Derivation.

$$f_X(x)\,\mathrm{d}x = f_X(\psi^{-1}(y)) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| \mathrm{d}y = f_Y(y)\,\mathrm{d}y$$

Linear transformations. A special case of diffeomorphic transformations. Consider $X \sim f_X(x)$ and $Y = aX + b$, for constants a, b $(a \neq 0)$. Then,

$$\begin{aligned} f_Y(y) &= \left| \frac{1}{a} \right| f_X\left(\frac{y-b}{a} \right) \\ \mathbb{E}[Y] &= a\mu_X + b \\ \text{Var}(Y) &= a^2 \sigma_X^2 \end{aligned}$$

Diffeomorphic transformation (n-D). Consider an r.v. $\boldsymbol{X} \sim f_X(\boldsymbol{x}), F_X(\boldsymbol{x})$ and another r.v. $\boldsymbol{Y} = \psi(\boldsymbol{X}) \sim f_Y(\boldsymbol{y}), F_Y(\boldsymbol{y})$ defined in terms of $\boldsymbol{X}$ by the diffeomorphism $\psi : \mathbb{R}^n \to \mathbb{R}^n$. Then, the density of $\boldsymbol{Y}$ is given by

$$f_Y(\boldsymbol{y}) = f_X(\psi^{-1}(\boldsymbol{y})) J_{\psi^{-1}}(\boldsymbol{y}),$$

where $J_{\psi^{-1}}(\boldsymbol{y})$ is the Jacobian:

$$J_{\psi^{-1}}(\boldsymbol{y}) \coloneqq \det \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_1}{\partial y_n} & \cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

General transformation ($\mathbb{R} \to \mathbb{R}$). Consider an r.v. $X \sim f_X(x), F_X(x)$ and another r.v. $Y = \psi(X) \sim f_Y(y), F_Y(y)$ defined in terms of X by the generic transformation $\psi : \mathbb{R} \to \mathbb{R}$. Then, the density of Y is given by

1. Defining the event of interest:

$$\mathbb{A}_\psi(y) \coloneqq \{Y \leq y\} = \{x \mid \psi(x) \leq y\}$$

2. Determining the cdf of Y:

$$F_Y(y) = \Pr(Y \leq y) = \Pr(X \in \mathbb{A}_\psi(y)) = \int_{\mathbb{A}_\psi(y)} f_X(x)\,\mathrm{d}x$$

3. Differentiating to get $f_Y(y)$:

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y)$$

General transformation ($\mathbb{R}^n \to \mathbb{R}^m$). Consider an r.v. $\boldsymbol{X} \sim f_X(\boldsymbol{x}), F_X(\boldsymbol{x})$ and another r.v. $\boldsymbol{Y} = \psi(\boldsymbol{X}) \sim f_Y(\boldsymbol{y}), F_Y(\boldsymbol{y})$ defined in terms of $\boldsymbol{X}$ by the generic transformation $\psi : \mathbb{R}^n \to \mathbb{R}^m$. Then, the density of $\boldsymbol{Y}$ is given by

1. Defining the event of interest:

$$\mathbb{A}_\psi(\boldsymbol{y}) \coloneqq \{\boldsymbol{Y} \leq \boldsymbol{y}\} = \{\boldsymbol{x} \mid \psi(\boldsymbol{x}) \leq \boldsymbol{y}\}$$

2. Determining the cdf of $\boldsymbol{Y}$:

$$F_Y(\boldsymbol{y}) = \Pr(\boldsymbol{Y} \leq \boldsymbol{y}) = \Pr(\boldsymbol{X} \in \mathbb{A}_\psi(\boldsymbol{y})) = \int \cdots \int_{\mathbb{A}_\psi(\boldsymbol{y})} f_X(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

3. Differentiating to get $f_Y(\boldsymbol{y})$:

$$f_Y(\boldsymbol{y}) = \frac{\partial}{\partial \boldsymbol{y}} F_Y(\boldsymbol{y})$$

Sum of independent r.v.s. Consider $X_1, \ldots, X_n \sim f(x_1, \ldots, x_n) = f_1(x_1) \times \cdots \times f_n(x_n)$ and $S_n = \sum_{i=1}^n X_i$. Then, the pdf $g_n(t)$ of S_n is given by the convolution

$$g_n(t) = f_1 * \cdots * f_n(t)$$

Proof. We begin with a simpler case of two random variables. $X_1, X_2 \sim f(x_1, x_2) = f_1(x_1) \times f_2(x_2)$. We want to determine the pdf $g_2(t)$ of $S_2 = X_1 + X_2$. We use the steps outlines in the generic general transformation above. The event of interest is, fixing t,

$A(t) = \{S_2 \leq t\}$. The pdf $g_2(t)$ is then

$$\begin{aligned} g_2(t) &= \frac{\mathrm{d}}{\mathrm{d}t} G_2(t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t} \int_{-\infty}^{\infty} \int_{-\infty}^{t-x_1} f_1(x_1) f_2(x_2) \,\mathrm{d}x_2 \,\mathrm{d}x_1 \\ &= \int_{-\infty}^{\infty} f_1(x_1) \left(\frac{\mathrm{d}}{\mathrm{d}t} \int_{-\infty}^{t-x_1} f_2(x_2) \,\mathrm{d}x_2 \right) \mathrm{d}x_1 \\ &= \int_{-\infty}^{\infty} f_1(x_1) f_2(t - x_1) \,\mathrm{d}x_1 \\ &= f_1 * f_2(t) \end{aligned}$$

The general case follows by induction: $S_n = S_{n-1} + X_n$. ■

Expectation of a transformed r.v. Let $\boldsymbol{X} \sim f_X(\boldsymbol{x})$ and $\psi : \mathbb{R}^n \to \mathbb{R}$ be a transformation of $\boldsymbol{X}$. Then, the expectation of the transformed r.v. $\psi(\boldsymbol{X})$ is given by

$$\mathbb{E}[\psi(\boldsymbol{X})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \psi(\boldsymbol{X}) f_X(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$$

Derivation. Let $Y = \psi(X) \sim f_Y(y)$.

$$\mathbb{E}[Y] \coloneqq \int_{-\infty}^{\infty} y f_Y(y) \,\mathrm{d}y = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \psi(\boldsymbol{x}) f_X(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \mathbb{E}[\psi(\boldsymbol{X})]$$

§2.6 The Normal Distribution

Def. **Univariate normal:** recall that $X \sim \mathcal{N}(\mu, \sigma^2)$ is distributed according to

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

and has mean μ and variance σ^2.

Def. **Bivariate normal:** $(X_1, X_2) \sim \phi(x_1, x_2; \rho)$ where $-1 < \rho < 1$ has pdf

$$\phi(x_1, x_2; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} (x_1^2 - 2\rho x_1 x_2 + x_2^2) \right)$$

Def. **Multivariate normal:** $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$, where

$$\boldsymbol{X} = (X_1, \ldots, X_n), \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \boldsymbol{C} = \mathrm{Cov}(\boldsymbol{X}) = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{n1} \\ \vdots & \ddots & \vdots \\ \sigma_{1n} & \cdots & \sigma_n^2 \end{bmatrix},$$

has the following properties:

- pdf
$$\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{C}) = \frac{1}{(2\pi)^{n/2}|C|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$
- Projections into lower dimensions are normal; marginals are normal.
- Uncorrelated (jointly) normal systems are independent: $\boldsymbol{C}$ is diagonal iff $X_1, \ldots, X_n$ are mutually independent.
- Linear transformations are normal

THM. **Sum of normals:** Let $X_1, \ldots, X_n$ be mutually independent where $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Then,
$$S_n = X_1 + \cdots + X_n \sim \mathcal{N}(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2)$$

Simplified proof. Let $X_1, X_2 \sim \mathcal{N}(0,1)$ and $S = X_1 + X_2$. We want to show that $S \sim (0,2)$. The pdf $g(t)$ of S is given by the convolution

$$\begin{aligned}
g(t) &= (\phi * \phi)(t) \\
&= \int_{-\infty}^{\infty} \phi(x)\phi(t-x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(t-x)^2}{2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{2x^2+t^2-2tx}{2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(\sqrt{2}x - \frac{t}{\sqrt{2}}\right)^2 + \frac{t^2}{2}}{2}}\,\mathrm{d}x && \text{(completing the square)} \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{4}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(\sqrt{2}x - \frac{t}{\sqrt{2}}\right)^2}{2}}\,\mathrm{d}x \\
&= \frac{1}{2\sqrt{\pi}} e^{-\frac{t^2}{4}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\,\mathrm{d}u && (u\text{-sub: } u = \sqrt{2}x - t/\sqrt{2}) \\
&= \frac{1}{2\sqrt{\pi}} e^{-\frac{t^2}{4}} && \text{(pdfs normalize to 1)}
\end{aligned}$$

Hence, we see that g is a normal pdf with mean $\mu = 0$ and variance $\sigma^2 = 2$. We conclude $S \sim \mathcal{N}(0,2)$. ■

§2.7 Other Distributions

DEF. **Cauchy:** an r.v. $X \sim Cauchy(\gamma)$ has the following properties

- pdf
$$c(x;\gamma) = \frac{1}{\pi\gamma\left[1+\left(\frac{x}{\gamma}\right)^2\right]}$$
- Expectation and variance are undefined
- Sum of Cauchys: for $X_1 \sim c(x_1;\gamma_1), X_2 \sim c(x_2;\gamma_2)$,
$$S = X_1 + X_2 \sim c(t;\gamma_1+\gamma_2)$$

DEF. **Gamma:** an r.v. $X \sim Gamma(n,\alpha)$ has pdf over the support $\mathbb{R}^+$

$$g(x) = \alpha\frac{(\alpha t)^{n-1}}{(n-1)!}e^{\alpha t}$$

Derivation from the convolution of exponentials. Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} Exponential(\alpha)$ and $S_n = X_1 + \cdots + X_n \sim g_n(t) = f^{*n}(t)$, $t > 0$. Fixing some $t > 0$,

- $n = 1$: $g_1(t) = \alpha e^{-\alpha x}$
- $n = 2$:
$$\begin{aligned} g_2(t) = (f*f)(t) = \int_{-\infty}^{\infty} f(x)f(t-x)\,\mathrm{d}x &= \int_0^t \alpha e^{-\alpha x}\alpha e^{-\alpha(t-x)}\,\mathrm{d}x \\ &= \int_0^t \alpha^2 e^{-\alpha t}\,\mathrm{d}x \\ &= \alpha^2 t e^{-\alpha t} \end{aligned}$$
- $n = 3$:
$$\begin{aligned} g_3(t) = (f*g_2)(t) = \int_{-\infty}^{\infty} g_2(x)f(t-x)\,\mathrm{d}x &= \int_0^t \alpha^2 x e^{-\alpha x}\alpha e^{-\alpha(t-x)}\,\mathrm{d}x \\ &= \int_0^t \alpha^3 x e^{-\alpha t}\,\mathrm{d}x \\ &= \alpha^3\frac{t^2}{2}e^{-\alpha t} \end{aligned}$$

- $n \geq 4$ by induction.

Note that $X \sim Gamma(1, \alpha) = Exponential(\alpha)$.

Poisson, redux.

Deriving Poisson from the exponential distribution. Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} Exponential(\alpha)$ be interarrival times. Let S_n be the time of the nth arrival and $N(t)$ be the number of arrivals up till time t.

From these definitions, we have the event $\{N(t) = n\} = \{S_n \leq t, S_{n+1} > t\}$. Note that $\{S_{n+1} > t\} = \{S_n > t\} \cup \{S_n \leq t, S_{n+1} > t\}$. Hence,

$$\begin{aligned}
\Pr(N(t) = n) &= \Pr(S_{n+1} > t) - \Pr(S_n > t) \\
&= e^{-\alpha t} \sum_{k=0}^{n} \frac{(\alpha t)^k}{k!} - e^{-\alpha t} \sum_{k=0}^{n-1} \frac{(\alpha t)^k}{k!} \\
&= e^{-\alpha t} \frac{(\alpha t)^n}{n!}, \quad n = 0, 1, 2, \ldots
\end{aligned}$$

Notice $N(t) \sim Poisson(\alpha t)$.

§2.8 Probabilistic Bounds

DEF. **Chebyshev's inequality:**

DEF. **Chernoff's inequality:**