Chapter II: Basics from probability theory and statistics

Information Retrieval & Data Mining
Universität des Saarlandes, Saarbrücken
Winter Semester 2011/12

Chapter II: Basics from Probability Theory and Statistics*

II.1 Probability Theory

Events, Probabilities, Random Variables, Distributions, Moment-Generating Functions, Deviation Bounds, Limit Theorems

Basics from Information Theory

II.2 Statistical Inference: Sampling and Estimation

Moment Estimation, Confidence Intervals

Parameter Estimation, Maximum Likelihood, EM Iteration

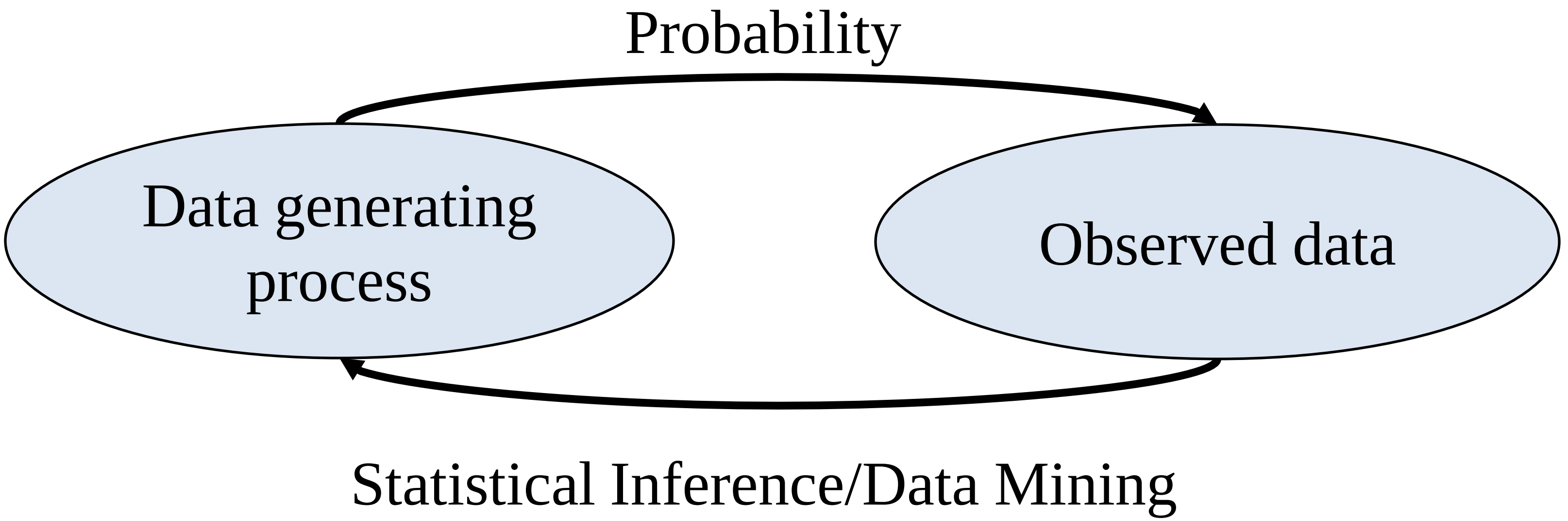
II.3 Statistical Inference: Hypothesis Testing and Regression

Statistical Tests, p-Values, Chi-Square Test

Linear and Logistic Regression

*mostly following L. Wasserman, with additions from other sources

II.1 Basic Probability Theory



- **Probability Theory**
  - Given a data generating process, what are the properties of the outcome?
- **Statistical Inference**
  - Given the outcome, what can we say about the process that generated the data?
  - How can we generalize these observations and make predictions about future outcomes?

Sample Spaces and Events

- A **sample space** Ω is a set of all possible outcomes of an experiment. (Elements e in Ω are called **sample outcomes** or **realizations**.)
- Subsets E of Ω are called **events**.

Example 1:

- If we toss a coin twice, then Ω = {HH, HT, TH, TT}.
- The event that the first toss is heads is A = {HH, HT}.

Example 2:

- Suppose we want to measure the temperature in a room.
- Let Ω = R = {-∞, ∞}, i.e., the set of the real numbers.
- The event that the temperature is between 0 and 23 degrees is A = [0, 23].

Probability

- A **probability space** is a triple (Ω, E, P) with
  - a sample space Ω of possible outcomes,
  - a set of events E over Ω,
  - and a **probability measure** $P: E \rightarrow [0,1]$.

Example: $P[\{HH, HT\}] = 1/2$; $P[\{HH, HT, TH, TT\}] = 1$

- **Three basic axioms of probability theory:**

  Axiom 1: $P[A] \geq 0$ (for any event A in E)

  Axiom 2: $P[\Omega] = 1$

  Axiom 3: If events $A_1, A_2, \ldots$ are disjoint, then $P[\cup_i A_i] = \sum_i P[A_i]$ (for countably many A_i).

Probability

<u>More properties</u> (derived from axioms)

$P[\emptyset] = 0$ (null/impossible event)

$P[\Omega] = 1$ (true/certain event, actually not derived but 2nd axiom)

$0 \le P[A] \le 1$

If $A \subseteq B$ then $P[A] \le P[B]$

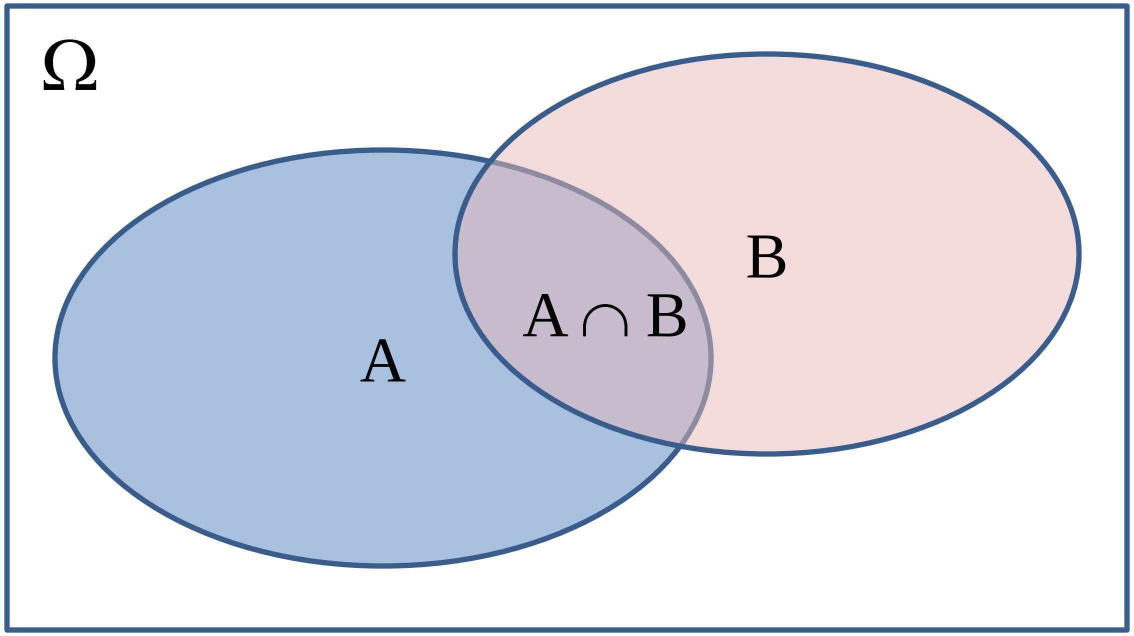
$P[A] + P[\neg A] = 1$

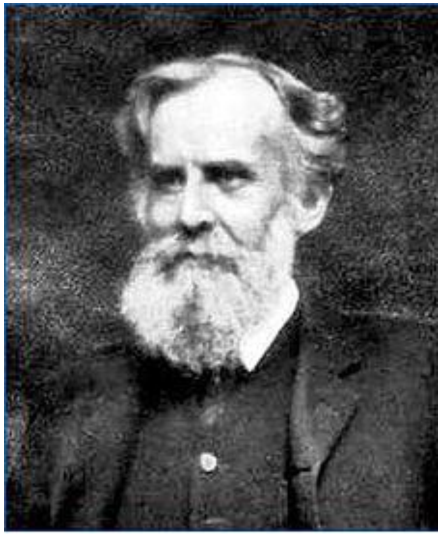
$P[A \cup B] = P[A] + P[B] - P[A \cap B]$ (inclusion-exclusion principle)

<u>Notes:</u>

- E is *closed* under $\cap$, $\cup$, and – with a countable number of operands (with finite Ω, usually $E=2^{\Omega}$).
- It is not always possible to assign a probability to every event in E if the sample space is large. Instead one may assign probabilities to a limited class of sets in E.

Venn Diagrams





John Venn
1834-1923

<u>Proof of the Inclusion-Exclusion Principle:</u>

$$
\begin{aligned}
P[A \cup B] &= P[\,(A \cap \neg B) \cup (A \cap B) \cup (\neg A \cap B)\,] \\
&= P[A \cap \neg B] + P[A \cap B] + P[\neg A \cap B] + P[A \cap B] - P[A \cap B] \\
&= P[(A \cap \neg B) \cup (A \cap B)] + P[(\neg A \cap B) \cup (A \cap B)] - P[A \cap B] \\
&= P[A] + P[B] - P[A \cap B]
\end{aligned}
$$

Independence and Conditional Probabilities

- Two **events** A, B of a probability space are **independent** if $P[A \cap B] = P[A]\, P[B]$.
- A finite **set of events** $A=\{A_1, \ldots, A_n\}$ is **independent** if for every subset $S \subseteq A$ the equation

$$P[\bigcap_{A_i \in S} A_i] = \prod_{A_i \in S} P[A_i]$$

holds.
- The **conditional probability** $P[A \mid B]$ of A under the condition (hypothesis) B is defined as:

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]}$$

- An event A is **conditionally independent** of B given C if $P[A \mid BC] = P[A \mid C]$.

Independence vs. Disjointness

Set-Complement

$$\boxed{P[\neg A] = 1 - P[A]}$$

Independence

$$\boxed{P[A \cap B] = P[A]\,P[B]}$$

$$\boxed{P[A \cup B] = 1 - (1 - P[A])(1 - P[B])}$$

Disjointness

$$\boxed{P[A \cap B] = 0}$$

$$\boxed{P[A \cup B] = P[A] + P[B]}$$

Identity

$$\boxed{P[A] = P[B] = P[A \cap B] = P[A \cup B]}$$

Murphy's Law

"Anything that can go wrong will go wrong."

Example:

- Assume a power plant has a probability of a failure on any given day of p.
- The plant may fail independently on any given day, i.e., the probability of a failure over n days is: **P[failure in n days] = 1 – (1 – p)^n**

Set p = 3 accidents / (365 days * 40 years) = 0.00021, then:

P[failure in 1 day] = 0.00021
P[failure in 10 days] = 0.002
P[failure in 100 days] = 0.020
P[failure in 1000 days] = 0.186
P[failure in 365*40 days] = 0.950

Birthday Paradox

In a group of n people, what is the probability that at least 2 people have the same birthday?

→ For n = 23, there is already a 50.7% probability of least 2 people having the same birthday.

Let N denote the event that in a group of n-1 people <u>a newly added person does not share a birthday with any other person</u>, then:

$P[N=1] = 365/365$, $P[N=2] = 364/365$, $P[N=3] = 363/365$, …

P[N'=n] = P[at least two birthdays in a group of n people coincide]

$$= 1 - P[N=1]\, P[N=2] \ldots P[N=n-1] = 1 - \prod_{k=1,\ldots,n-1} (1 - k/365)$$

$P[N'=1] = 0$

$P[N'=10] = 0.117$

$P[N'=23] = 0.507$

$P[N'=41] = 0.903$

$P[N'=366] = 1.0$

Total Probability and Bayes' Theorem

The Law of Total Probability:

For a partitioning of Ω into events $A_1, \ldots, A_n$:

$$P[B] = \sum_{i=1}^{n} P[B \mid A_i]\, P[A_i]$$

Thomas Bayes
1701-1761

Bayes' Theorem: $$P[A \mid B] = \frac{P[B \mid A]\, P[A]}{P[B]}$$

P[A|B] is called ***posterior probability***

P[A] is called ***prior probability***

Random Variables

How to link sample spaces and events to actual data / observations?

Example:

Let's flip a coin twice, and let X denote the number of heads we observe. Then what are the probabilities P[X=0], P[X=1], etc.?

$P[X=0] = P[\{TT\}] = 1/4$

$P[X=1] = P[\{HT, TH\}] = 1/4 + 1/4 = 1/2$

$P[X=2] = P[\{HH\}] = 1/4$

x	**P(X=x)**
0	1/4
1	1/2
2	1/4

Distribution of X

What is the probability of P[X=3] ?

Random Variables

- A **random variable (RV)** X on the probability space (Ω, E, P) is a function $X: \Omega \rightarrow M$ with $M \subseteq R$ s.t. $\{e \mid X(e) \leq x\} \in E$ for all $x \in M$ (X is observable).

Example: (**Discrete RV**)
Let's flip a coin 10 times, and let X denote the number of heads we observe.
If e = HHHHHTHHTT, then X(e) = 7.

Example: (**Continuous RV**)
Let's flip a coin 10 times, and let X denote the ratio between heads and tails we observe. If e = HHHHHTHHTT, then X(e) = 7/3.

Example: (**Boolean RV**, special case of a discrete RV)
Let's flip a coin twice, and let X denote the event that heads occurs first.
Then X=1 for {HH, HT}, and X=0 otherwise.

Distribution and Density Functions

- $F_X: M \rightarrow [0,1]$ with $F_X(x) = P[X \leq x]$ is the **cumulative distribution function (cdf)** of X.

- For a countable set M, the function $f_X: M \rightarrow [0,1]$ with $f_X(x) = P[X = x]$ is called the **probability density function (pdf)** of X; in general $f_X(x)$ is $F'_X(x)$.

- For a random variable X with distribution function F, the inverse function $F^{-1}(q) := \inf\{x \mid F(x) > q\}$ for $q \in [0,1]$ is called **quantile function** of X.

(the 0.5 quantile (aka. "50^{th} percentile") is called **median**)

> Random variables with countable M are called ***discrete***, otherwise they are called ***continuous***.
> For discrete random variables, the density function is also referred to as the ***probability mass function***.

Important Discrete Distributions

- **Uniform** distribution over {1, 2, ..., m}:

$$P[X=k] = f_X(k) = \frac{1}{m} \quad for\ 1 \le k \le m$$

- **Bernoulli** distribution (single coin toss with parameter p; X: head or tail):

$$P[X=k] = f_X(k) = p^k (1-p)^{1-k} \quad for\ k \in \{0,1\}$$

- **Binomial** distribution (coin toss n times repeated; X: #heads):

$$P[X=k] = f_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad for\ k \le n$$

- **Geometric** distribution (X: #coin tosses until first head):

$$P[X=k] = f_X(k) = (1-p)^k\ p$$

- **Poisson** distribution (with rate λ):

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

- **2-Poisson mixture** (with $a_1+a_2=1$):

$$P[X=k] = f_X(k) = a_1 e^{-\lambda_1} \frac{\lambda_1^k}{k!} + a_2 e^{-\lambda_2} \frac{\lambda_2^k}{k!}$$

Important Continuous Distributions

- **Uniform** distribution in the interval [a,b]

$$f_X(x) = \frac{1}{b-a} \quad for\ a \le x \le b\ \ (0\ otherwise)$$

- **Exponential** distribution (e.g. time until next event of a Poisson process) with rate $\lambda = \lim_{\Delta t \to 0}$ (# events in Δt) / Δt :

$$f_X(x) = \lambda e^{-\lambda x} \quad for\ x \ge 0\ (0\ otherwise)$$

- **Hyper-exponential** distribution:

$$f_X(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}$$

- **Pareto** distribution:
  Example of a “heavy-tailed” distribution with

$$f_X(x) \to \frac{a}{b}\left(\frac{b}{x}\right)^{a+1} \quad for\ x > b,\ 0\ otherwise$$

- **Logistic** distribution:

$$F_X(x) = \frac{1}{1+e^{-x}} \qquad f_X(x) \to \frac{c}{x^{\alpha+1}}$$

Normal (Gaussian) Distribution

- **Normal distribution N(μ,σ²)** (Gauss distribution; approximates sums of independent, identically distributed random variables): $$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- **Normal (cumulative) distribution function N(0,1):**

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx$$

Carl Friedrich Gauss, 1777-1855

<u>Theorem:</u>

Let X be Normal distributed with expectation μ and variance σ².

Then $Y := \frac{X - \mu}{\sigma}$

is Normal distributed with expectation 0 and variance 1.

Multidimensional (Multivariate) Distributions

Let $X_1, \ldots, X_m$ be random variables over the same probability space with domains $dom(X_1), \ldots, dom(X_m)$.

The **joint distribution** of $X_1, \ldots, X_m$ has the density function $f_{X_1,\ldots,X_m}(x_1,\ldots,x_m)$

$$with \sum_{x_1 \in dom(X_1)} \ldots \sum_{x_m \in dom(X_m)} f_{X_1,\ldots,X_m}(x_1,\ldots,x_m) = 1 \quad \text{(discrete case)}$$

$$or \int_{dom(X_1)} \ldots \int_{dom(X_m)} f_{X_1,\ldots,X_m}(x_1,\ldots,x_m)\, dx_m \ldots dx_1 = 1 \quad \text{(continuous case)}$$

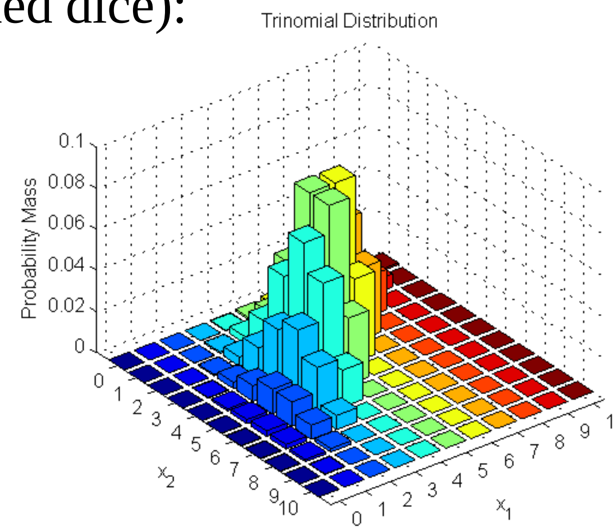
The **marginal distribution** of X_i in the joint distribution of $X_1, \ldots, X_m$ has the density function

$$\sum_{x_1} \ldots \sum_{x_{i-1}} \sum_{x_{i+1}} \ldots \sum_{x_m} f_{X_1,\ldots,X_m}(x_1,\ldots,x_m) \;\; or \quad \text{(discrete case)}$$

$$\int_{X_1} \ldots \int_{X_{i-1}} \int_{X_{i+1}} \ldots \int_{X_m} f_{X_1,\ldots,X_m}(x_1,\ldots,x_m)\; dx_m \;\ldots\; dx_{i+1}\; dx_{i-1} \ldots dx_1 \quad \text{(continuous case)}$$

Important Multivariate Distributions

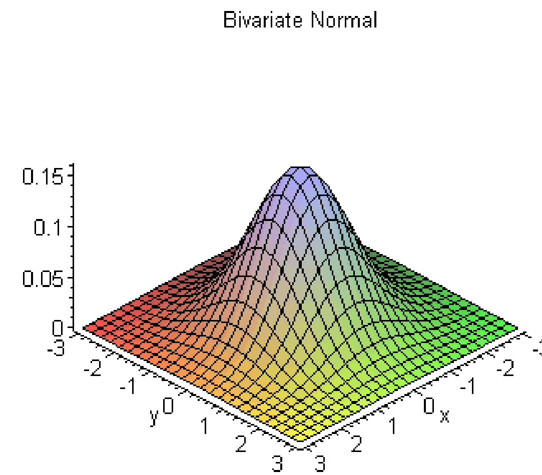
Multinomial distribution (n, m) (n trials with m-sided dice):



$$P[X_1 = k_1 \wedge \ldots \wedge X_m = k_m] =$$

$$f_{X_1,\ldots,X_m}(k_1,\ldots,k_m) = \binom{n}{k_1 \ldots k_m} p_1^{k_1} \ldots p_m^{k_m}$$

$$with \binom{n}{k_1 \ldots k_m} := \frac{n!}{k_1! \ldots k_m!}$$



Multidimensional Gaussian distribution ($\vec{\mu}$, Σ):

$$f_{X_1,\ldots,X_m}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}$$

with covariance matrix Σ with Σ_{ij} := Cov(X_i,X_j)

(Plots from http://www.mathworks.de/)

Expectation Values, Moments & Variance

For a discrete random variable X with density f_X

$$E[X] = \sum_{k \in M} k \; f_X(k)$$ is the **expectation value (mean)** of X

$$E[X^i] = \sum_{k \in M} k^i \; f_X(k)$$ is the **i-th moment** of X

$$V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$ is the **variance** of X

For a continuous random variable X with density f_X

$$E[X] = \int_{-\infty}^{+\infty} x \; f_X(x)\, dx$$ is the **expectation value (mean)** of X

$$E[X^i] = \int_{-\infty}^{+\infty} x^i \; f_X(x)\, dx$$ is the **i-th moment** of X

$$V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$ is the **variance** of X

Theorem: Expectation values are additive: $E[X+Y] = E[X] + E[Y]$
(distributions generally not)

Properties of Expectation and Variance

- $\mathbf{E[aX+b] = aE[X]+b}$ for constants a, b
- $\mathbf{E[X_1+X_2+...+X_n] = E[X_1] + E[X_2] + ... + E[X_n]}$
  (i.e. expectation values are generally additive, but distributions are not!)
- $\mathbf{E[XY] = E[X]E[Y]}$ if X and Y are independent
- $\mathbf{E[X_1+X_2+...+X_N] = E[N]\,E[X]}$
  if $X_1, X_2, ..., X_N$ are independent and identically distributed **(iid)** RVs with mean E[X] and N is a stopping-time RV
- $\mathbf{Var[aX+b] = a^2\,Var[X]}$ for constants a, b
- $\mathbf{Var[X_1+X_2+...+X_n] = Var[X_1] + Var[X_2] + ... + Var[X_n]}$
  if $X_1, X_2, ..., X_n$ are independent RVs
- $\mathbf{Var[X_1+X_2+...+X_N] = E[N]\,Var[X] + E[X]^2\,Var[N]}$
  if $X_1, X_2, ..., X_N$ are iid RVs with mean E[X] and variance Var[X] and N is a stopping-time RV

Correlation of Random Variables

Covariance of random variables X_i and X_j

$$Cov(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])]$$

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

Correlation coefficient of X_i and X_j

$$\rho(X_i, X_j) := \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)}\sqrt{Var(X_j)}}$$

Conditional expectation of X given Y=y

$$E[X \mid Y = y] = \begin{cases} \sum x\, f_{X|Y}(x \mid y) & \text{(discrete case)} \\ \int x\, f_{X|Y}(x \mid y)\, dx & \text{(continuous case)} \end{cases}$$

Transformations of Random Variables

Consider expressions r(X,Y) over RVs, such as X+Y, max(X,Y), etc.

1. For each z find $A_z = \{(x,y) \mid r(x,y) \le z\}$
2. Find cdf $F_Z(z) = P[r(x,y) \le z] = \int\int_{A_Z} f_{X,Y}(x,y)\,dx\,dy$
3. Find pdf $f_Z(z) = F'_Z(z)$

Important case: **Sum of independent RVs** (non-negative) Z = X+Y

$$F_Z(z) = P[r(x,y) \le z] = \int_y\int_x {}_{x+y\le z} f_X(x) f_Y(y)\,dx\,dy$$

"Convolution"

$$= \int_{y=0}^{z-x}\int_{x=0}^{z} f_X(x) f_Y(y)\,dx\,dy$$

$$= \int_{x=0}^{z} f_X(x) F_Y(z-x)\,dx$$

Discrete case:

$$F_Z(z) = \sum_x \sum_y {}_{x+y\le z} f_X(x) f_Y(y)$$

$$= \sum_{x=0}^{z} f_X(x) F_Y(z-x)$$

Generating Functions and Transforms

X, Y, ...: continuous random variables with non-negative real values

$$M_X(s) = \int_0^\infty e^{sx} f_X(x)dx = E[e^{sX}]:$$

Moment-generating function of X

$$f^*_X(s) = \int_0^\infty e^{-sx} f_X(x)dx = E[e^{-sX}]$$

Laplace-Stieltjes transform (LST) of X

A, B, ...: discrete random variables with non-negative integer values

$$G_A(z) = \sum_{i=0}^{\infty} z^i f_A(i) = E[z^A]:$$

Generating function of A (z transform)

$$f_A^*(-s) = M_A(s) = G_A(e^s)$$

Laplace-Stieltjes transform of A

Examples:

Exponential:

$$f_X(x) = \alpha e^{-\alpha x}$$

$$f^*_X(s) = \frac{\alpha}{\alpha + s}$$

Erlang-k:

$$f_X(x) = \frac{\alpha k(\alpha k x)^{k-1}}{(k-1)!} e^{-\alpha k x}$$

$$f^*_X(s) = \left(\frac{k\alpha}{k\alpha + s}\right)^k$$

Poisson:

$$f_A(k) = e^{-\alpha} \frac{\alpha^k}{k!}$$

$$G_A(z) = e^{\alpha(z-1)}$$

Properties of Transforms

Convolution of independent random variables:

$$F_{X+Y}(z) = \int_0^z f_X(x)\, F_Y(z-x)\, dx$$

$$M_{X+Y}(s) = M_X(s) M_Y(s)$$

$$f^*_{X+Y}(s) = f^*_X(s)\, f^*_Y(s)$$

(continuous case)

$$F_{A+B}(k) = \sum_{i=0}^{k} f_A(i)\, F_B(k-i)$$

$$G_{A+B}(z) = G_A(z) G_B(z)$$

(discrete case)

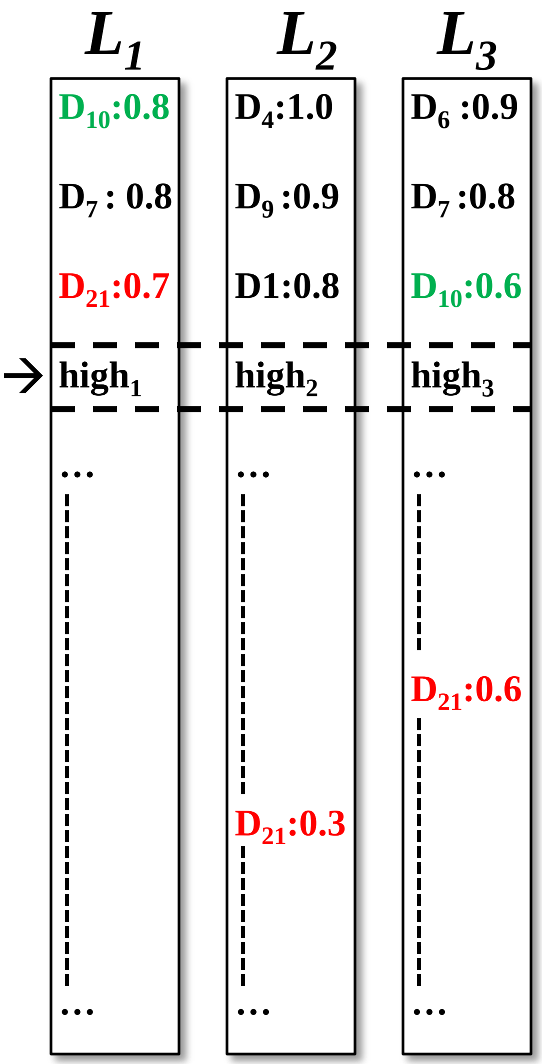
Many more properties for other transforms, see, e.g.:

L. Wasserman: All of Statistics

Arnold O. Allen: Probability, Statistics, and Queueing Theory

Use Case: Score prediction for fast Top-k Queries

[Theobald, Schenkel, Weikum: VLDB'04]



Given: Inverted lists L_i with continuous score distributions captured by independent RV's S_i

Want to predict: $P\left[\sum_i S_i > \delta\right]$

- Consider score intervals $[0, high_i]$ at current scan positions in L_i, then $f_i(x) = 1/high_i$ (assuming uniform score distributions)

- Convolution S_1+S_2 is given by

$$F_{S_1+S_2}(z) = \int_0^z f_{S_1}(x)\, F_{S_2}(z-x)\, dx$$

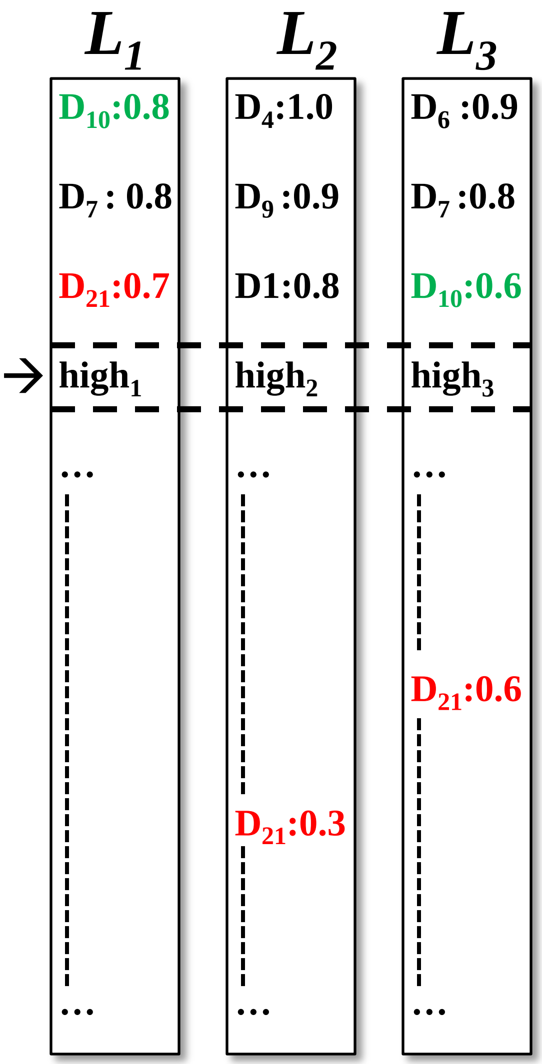
- But each factor is non-zero in $0 \le x \le high_1$ and $0 \le z\text{-}x \le high_2$ only (for $high_1 \le high_2$), thus

$$f(x) = \begin{cases} x/(high_1 \cdot high_2) & for\ 0 \le x \le high_1 \\ 1/high_2 & for\ high_1 < x \le high_2 \\ 1/high_1 + 1/high_2 - x/(high_1 \cdot high_2) & \\ & for\ high_2 < x \le high_1 + high_2 \end{cases}$$

→ Cumbersome amount of case differentiations

Use Case: Score prediction for fast Top-k Queries

[Theobald, Schenkel, Weikum: VLDB'04]



Given: Inverted lists L_i with continuous score distributions captured by independent RV's S_i

Want to predict: $P\left[\sum_i S_i > \delta\right]$

- **Instead:** Consider the moment-generating function for each S_i

$$M_i(s) = \int_0^s e^{sx} f_i(x)\,dx = E\left[e^{sS_i}\right]$$

- For independent S_i, the moment of the convolution over all S_i is given by

$$M(s) = \prod_i M_i(s)$$

- Apply ***Chernoff-Hoeffding bound*** on tail distribution

$$P\left[\sum_i S_i > \delta\right] \le \inf_{s \ge 0}\{e^{-s\delta} M(s)\}$$

→ Prune D_{21} if $P[S_2+S_3 > \delta] \le \varepsilon$ (using $\delta = 1.4\text{-}0.7$ and a small confidence threshold for ε, e.g., $\varepsilon = 0.05$)

Inequalities and Tail Bounds

Markov inequality: $P[X \geq t] \leq E[X] / t$ for $t > 0$ and non-neg. RV X

Chebyshev inequality: $P[\, |X-E[X]| \geq t] \leq Var[X] / t^2$ for $t > 0$ and non-neg. RV X

Chernoff-Hoeffding bound: $P[X \geq t] \leq \inf \left\{ e^{-\theta t} M_X(\theta) \mid \theta \geq 0 \right\}$

Corollary: $$P\left[\left|\frac{1}{n}\sum X_i - p\right| \geq t\right] \leq 2e^{-2nt^2}$$ for Bernoulli(p) iid. RVs $X_1, \ldots, X_n$ and any $t > 0$

Mill's inequality: $$P\left[|Z| > t\right] \leq \frac{\sqrt{2}}{\pi} \frac{e^{-t^2/2}}{t}$$ for N(0,1) distr. RV Z and $t > 0$

Cauchy-Schwarz inequality: $$E[XY] \leq \sqrt{E[X^2]E[Y^2]}$$

Jensen's inequality: $E[g(X)] \geq g(E[X])$ for convex function g

$E[g(X)] \leq g(E[X])$ for concave function g

(g is convex if for all $c \in [0,1]$ and x_1, x_2: $g(cx_1 + (1-c)x_2) \leq cg(x_1) + (1-c)g(x_2)$)

Convergence of Random Variables

Let $X_1, X_2, \ldots$ be **a sequence of RVs** with cdf's $F_1, F_2, \ldots$, and let X be another RV with cdf F.

- X_n **converges** to X **in probability**, $X_n \to_P X$, if for every $\varepsilon > 0$
  $P[|X_n - X| > \varepsilon] \to 0$ as $n \to \infty$
- X_n **converges** to X **in distribution**, $X_n \to_D X$, if
  $\lim_{n\to\infty} F_n(x) = F(x)$ at all x for which F is continuous
- X_n **converges** to X **in quadratic mean**, $X_n \to_{qm} X$, if
  $E[(X_n - X)^2] \to 0$ as $n \to \infty$
- X_n **converges** to X **almost surely**, $X_n \to_{as} X$, if $P[X_n \to X] = 1$

Weak law of large numbers (for $\bar{X}_n = \sum_{i=1..n} X_i / n$)
if $X_1, X_2, \ldots, X_n, \ldots$ are iid RVs with mean E[X], then $\bar{X}_n \to_P E[X]$
that is: $\lim_{n\to\infty} P[|\bar{X}_n - E[X]| > \varepsilon] = 0$

Strong law of large numbers:
if $X_1, X_2, \ldots, X_n, \ldots$ are iid RVs with mean E[X], then $\bar{X}_n \to_{as} E[X]$
that is: $P[\lim_{n\to\infty} |\bar{X}_n - E[X]| > \varepsilon] = 0$

Convergence & Approximations

Theorem: (Binomial converges to Poisson)
Let X be a random variable with Binomial distribution with parameters n and p := λ/n with large n and small constant λ << 1.

Then $\lim_{n\to\infty} f_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$

Theorem: (Moivre-Laplace: Binomial converges to Gaussian)
Let X be a random variable with Binomial distribution with parameters n and p. For -∞ < a ≤ b < ∞ it holds that:

$$\lim_{n\to\infty} P[a \le \frac{X - np}{\sqrt{np(1-p)}} \le b] = \Phi(b) - \Phi(a)$$

Φ(z) is the Normal distribution function N(0,1); a, b are integers

Central Limit Theorem

Theorem:

Let $X_1, \ldots, X_n$ be n independent, identically distributed (iid) random variables with expectation μ and variance σ^2. The distribution function F_n of the random variable $Z_n := X_1 + \ldots + X_n$ converges to a Normal distribution $N(n\mu, n\sigma^2)$ with expectation nμ and variance $n\sigma^2$. That is, for $-\infty < x \le y < \infty$ it holds that:

$$\lim_{n\to\infty} P[x \le \frac{Z_n - n\mu}{\sqrt{n}\,\sigma} \le y] = \Phi(y) - \Phi(x)$$

Corollary:

$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ converges to a Normal distribution $N(\mu, \sigma^2/n)$ with expectation μ and variance σ^2/n .

Elementary Information Theory

Let f(x) be the probability (or relative frequency) of the x-th symbol in some text d. The **entropy** of the text (or the underlying prob. distribution f) is: $H(d) = \sum_x f(x) \log_2 \frac{1}{f(x)}$

H(d) is a lower bound for the *bits per symbol* needed with optimal coding (compression).

For two prob. distributions f(x) and g(x) the **relative entropy (Kullback-Leibler divergence)** of f to g is:

$$D(f \| g) := \sum_x f(x) \log_2 \frac{f(x)}{g(x)}$$

Relative entropy is a measure for the (dis-)similarity of two probability or frequency distributions. It corresponds to the *average number of additional bits* needed for coding information (events) with distribution f when using an optimal code for distribution g.

The **cross entropy** of f(x) to g(x) is:

$$H(f,g) := H(f) + D(f \| g) = -\sum_x f(x) \log g(x)$$

Compression

- Text is sequence of symbols (with specific frequencies)
- Symbols can be
  - letters or other characters from some alphabet Σ
  - strings of fixed length (e.g. trigrams, "shingles")
  - or words, bits, syllables, phrases, etc.

Limits of compression:

Let p_i be the probability (or relative frequency) of the i-th symbol in text d

Then the ***entropy*** of the text: $H(d) = \sum_i p_i \log_2 \frac{1}{p_i}$

is a ***lower bound*** for the average number of bits per symbol in any compression (e.g. Huffman codes)

Note:

Compression schemes such as *Ziv-Lempel* (used in zip) are better because they consider context beyond single symbols; with appropriately generalized notions of entropy, the lower-bound theorem does still hold.

Summary of Section II.1

- **Bayes' Theorem**: very simple, very powerful
- **RVs** as a fundamental, sometimes subtle concept
- Rich variety of well-studied **distribution functions**
- **Moments** and **moment-generating functions** capture distributions
- **Tail bounds** useful for non-tractable distributions
- **Normal** distribution: limit of sum of iid RVs
- **Entropy** measures (incl. **KL divergence**)

  capture complexity and similarity of prob. distributions

Reference Tables on Probability Distributions and Statistics (1)

Appendix A

Statistical Tables

A.1 Discrete Random Variables

Table 1A. Properties of Some Common Discrete Random Variables[1]

Random Variable	Parameters	$p(\cdot)$
Bernoulli	$0 < p < 1$	$p(k) = p^k q^{1-k}$ $k = 0, 1$
Binomial	n $0 < p < 1$	$p(k) = \binom{n}{k} p^k q^{n-k}$, $k = 0, 1, \ldots, n$
Multinomial	n, r, p_i, k_i $\sum_{i=1}^{r} p_i = 1$ $\sum_{i=1}^{r} k_i = n$,	$p(\overline{k}) = \frac{n!}{k_1! k_2! \cdots k_r!} p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r}$ where $\overline{k} = (k_1, k_2, \ldots, k_r)$

[1] $q = 1 - p$.

Table 1A. (continued)

Random Variable	Parameters	$p(\cdot)$
Hypergeometric	$N > 0$ $n, k \geq 0$	$p(k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}$, $k = 0, 1, \ldots, n$, where $k \leq r$ and $n - k \leq N - r$.
Multivariate Hypergeometric	$\sum_{i=1}^{l} r_i = N$	$p(k_1, k_2, \ldots, k_l) = \frac{\binom{r_1}{k_1}\binom{r_2}{k_2}\cdots\binom{r_l}{k_l}}{\binom{N}{n}}$, for $k_i \in \{0, 1, \ldots, n\}$, $k_i \leq r_i \ \forall i$ and $\sum_{i=1}^{l} k_i = n$.
Geometric	$0 < p < 1$	$p(k) = q^k p, \quad k = 0, 1, \ldots$.
Pascal (negative binomial)	$0 < p < 1$ r positive integer	$p(k) = \binom{r+k-1}{k} p^r q^k$, $k = 0, 1, \cdots$
Poisson	$\alpha > 0$	$p(k) = e^{-\alpha} \frac{\alpha^k}{k!}, \quad k = 0, 1, \cdots$

Source: Arnold O. Allen, Probability, Statistics, and Queueing Theory with Computer Science Applications, Academic Press, 1990

Reference Tables on Probability Distributions and Statistics (2)

Table 1B. Properties of Some Common Discrete Random Variables[2]

Random Variable	z-transform $g[z]$	$E[X]$	$\mathrm{Var}[X]$
Bernoulli	$q+pz$	p	pq
Binomial	$(q+pz)^n$	np	npq
Multinomial	$(p_1z_1+p_2z_2 + \cdots + p_rz_r)^n$	$E[X_i] = np_i$	$\mathrm{Var}[X_i] = np_iq_i$
Hypergeometric	—	$\frac{nr}{N}$	$\frac{nr(N-r)(N-n)}{N^2(N-1)}$
Multivariate Hypergeometric	—	—	—
Geometric	$\frac{p}{1-qz}$	$\frac{q}{p}$	$\frac{q}{p^2}$
Pascal (negative binomial)	$p^r(1-qz)^{-r}$	$\frac{rq}{p}$	$\frac{rq}{p^2}$
Poisson	$e^{\alpha(z-1)}$	α	α

[2] $q_i = 1 - p_i$.

Table 2A. Properties of Some Common Continuous Random Variables

Random Variable	Parameters	Density $f(\cdot)$
Uniform	$a<b$	$\frac{1}{b-a}$, $a \le x \le b$, 0 otherwise
Exponential	$\alpha > 0$	$f(x) = \alpha e^{-\alpha x}$, $x>0$, 0 if $x \le 0$
Gamma	$\beta, \alpha > 0$	$f(x) = \frac{\alpha(\alpha x)^{\beta-1}}{\Gamma(\beta)} e^{-\alpha x}$, $x > 0$ 0, $x \le 0$
Erlang-k	$k > 0$ $\mu > 0$	$f(x) = \frac{\mu k(\mu k x)^{k-1}}{(k-1)!} e^{-\mu k x}$, $x>0$ 0, $x \le 0$
H_k [3]	$q_i, \mu_i > 0$ $\sum_{i=1}^{k} \frac{q_i}{\mu_i} = \frac{1}{\mu}$	$f(x) = \sum_{i=1}^{k} q_i \mu_i e^{-\mu_i x}$, $x > 0$ 0, $x \le 0$
Chi-square	$n > 0$	$f(x) = \frac{x^{((n/2)-1)} e^{-x/2}}{2^{n/2}\Gamma(n/2)}$, $x > 0$, 0 if $x \le 0$
Normal	$\sigma > 0$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$
Student's t	n	$f(x) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\Gamma(n/2)} \left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$
F	n, m	$f(x) = \frac{(n/m)^{n/2}\Gamma[(n+m)/2] x^{((n/2)-1)}}{\Gamma(n/2)\Gamma(m/2)(1+(n/m)x)^{(n+m)/2}}$, $x > 0$

[3] Hyperexponential with k stages.

Source: Arnold O. Allen, Probability, Statistics, and Queueing Theory with Computer Science Applications, Academic Press, 1990

Reference Tables on Probability Distributions and Statistics (3)

Table 2B. Properties of Some Common Continuous Random Variables

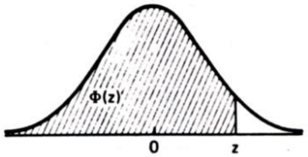
Random Variable	$E[X]$	$\mathrm{Var}[X]$	Laplace–Stieltjes Transform $X^*[\theta]$
Uniform	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{-b\theta}-e^{-a\theta}}{\theta(a-b)}$
Exponential	$\frac{1}{\alpha}$	$\frac{1}{\alpha^2}$	$\frac{\alpha}{\alpha+\theta}$
Gamma	$\frac{\beta}{\alpha}$	$\frac{\beta}{\alpha^2}$	$\left(\frac{\alpha}{\alpha+\theta}\right)^{\beta}$
Erlang-k	$\frac{1}{\mu}$	$\frac{1}{k\mu^2}$	$\left(\frac{k\mu}{k\mu+\theta}\right)^{k}$
H_k [4]	$\frac{1}{\mu}$	$\left(2\sum_{i=1}^{k}\frac{q_i}{\mu_i^2}\right)-\frac{1}{\mu^2}$	$\sum_{i=1}^{k}\frac{q_i\mu_i}{\mu_i+\theta}$
Chi-square	n	$2n$	$\left(\frac{1}{1+2\theta}\right)^{n/2}$
Normal	μ	σ^2	$\exp\left(-\theta\mu-\frac{1}{2}\theta^2\sigma^2\right)$
Student's t	0 for $n>1$	$\frac{n}{n-2}$ for $n>2$	does not exist
F	$\frac{m}{m-2}$ if $m>2$	$\frac{m^2(2n+2m-4)}{n(m-2)^2(m-4)}$ if $m>4$	does not exist

[4] Hyperexponential with k stages.

A.3 Statistical Tables

Table 3

The Normal Distribution Functions $\Phi(z) = \int_{-\infty}^{z} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,dt$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995

Source: Arnold O. Allen, Probability, Statistics, and Queueing Theory with Computer Science Applications, Academic Press, 1990

Reference Tables on Probability Distributions and Statistics (4)

A.4 The Laplace–Stieltjes Transform

Table 10. Laplace Transform Properties and Identities[5]

	Function	Transform
1.	$f(t)$	$f^*[\theta] = \int_{0-}^{\infty} e^{-\theta t} f(t)\,dt$
2.	$af(t) + bg(t)$	$af^*[\theta] + bg^*[\theta]$
3.	$f\left(\frac{t}{a}\right), \quad a > 0$	$af^*[a\theta]$
4.	$f(t-a)$ for $t \geq a$	$e^{-a\theta} f^*[\theta]$
5.	$e^{-at} f(t)$	$f^*[\theta + a]$
6.	$tf(t)$	$-\frac{df^*[\theta]}{d\theta}$
7.	$t^n f(t)$	$(-1)^n \frac{d^n f^*[\theta]}{d\theta^n}$
8.	$\int_0^t f(u) g(t-u)\,du$	$f^*[\theta] g^*[\theta]$
9.	$\frac{df(t)}{dt}$	$\theta f^*[\theta] - f(0)$
10.	$\frac{d^n f(t)}{dt^n}$	$\theta^n f^*[\theta] - \sum_{i=1}^{n} \theta^{n-i} f^{(i-1)}(0)$
11.	$\int_0^t f(x)\,dx$	$\frac{f^*[\theta]}{\theta}$
12.	$\frac{\partial f(t)}{\partial a}$ a a parameter	$\frac{\partial f^*[\theta]}{\partial a}$

[5]All functions f are assumed to be piecewise continuous and of exponential order. That is, there exist positive constants M and a such that $|f(t)| \leq Me^{at}$ for $t \geq 0$.

Table 11. Laplace Transform Pairs

	Function	Transform
1.	$f(t)$	$f^*[\theta] = \int_{0-}^{\infty} e^{-\theta t} f(t)\,dt$
2.	$f(t) = c$	$\frac{c}{\theta}$
3.	$t^n, \quad n = 1, 2, 3, \cdots$	$\frac{n!}{\theta^{n+1}}$
4.	$t^a, \quad a > 0$	$\frac{\Gamma(a+1)}{\theta^{a+1}}$
5.	e^{at}	$\frac{1}{\theta - a}, \quad \theta > a$
6.	te^{at}	$\frac{1}{(\theta - a)^2}, \quad \theta > a$
7.	$t^n e^{at}$	$\frac{n!}{(\theta - a)^{n+1}}, \quad \theta > a$
8.[6]	$\delta(t)$	1
9.	$\delta(t-a)$	$e^{-a\theta}$
10.[7]	$U(t-a)$	$\frac{e^{-a\theta}}{\theta}$
11.	$f(t-a)U(t-a)$	$e^{-a\theta} f^*[\theta]$

[6]The Dirac delta function $\delta(\cdot)$ is defined by $\delta(t) = 0$ for $t \neq 0$ but $\int_{a-\epsilon}^{a+\epsilon} \delta(t-a) f(t)\,dt = f(a)$ for each f and each $\epsilon > 0$.

[7]The unit step function $U(\cdot)$ is defined by

$$U(t-a) = \begin{cases} 0 & \text{for } t < a \\ 1 & \text{for } t \geq a. \end{cases}$$

Source: Arnold O. Allen, Probability, Statistics, and Queueing Theory with Computer Science Applications, Academic Press, 1990