Condensed Probability Theory and Statistics

Arranged by Curtis Toupin

A collection of common and useful results known in the field of probability and statistics

Curtis Toupin, Ottawa, Canada, 2020

Remark

From time to time, there will be remarks that contain vital information for the reader. When such a remark arises, it will be contained in a box like this one.

CONTENTS

Part I

Probability Theory

SECTION 1

OVERVIEW

1.1 Probability Spaces

Definition 1.1 A *σ-algebra* on a set Ω is a collection, $\mathcal{F}$, of subsets of Ω satisfying the following:

- $\Omega \in \mathcal{F}$
- $\mathcal{F}$ is closed under complement. That is, for all $A \in \mathcal{F}$, $A' \in \mathcal{F}$ as well, and
- $\mathcal{F}$ is closed under countable unions. That is, for any collection of sets, $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$, we have

  $$\bigcup_{n\in\mathbb{N}} A_i \in \mathcal{F}$$

  as well.

Proposition 1.2 For any σ-algebra $\mathcal{F}$ over a set Ω,

- $\emptyset \in \mathcal{F}$
- $\mathcal{F}$ is closed under countable intersection

Definition 1.3 Let $\mathcal{F}$ be a σ-algebra over a set Ω. A *measure* is a function $\mu : \mathcal{F} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, the extended real numbers satisfying the following:

- for all $A \in \mathcal{F}$, $\mu(A) \geq 0$
- $\mu(\emptyset) = 0$, and
- For any pairwise disjoint collection of sets, $\{A_i\}_{i\in\mathbb{N}}$ of pairwise disjoint sets in $\mathcal{F}$,

$$\mu\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mu(A_i).$$

The pair $(\Omega, \mathcal{F})$ is called a *measurable space*. Members of $\mathcal{F}$ are called *measurable sets*. The triplet $(\Omega, \mathcal{F}, \mu)$ is known as a *measure space*.

Definition 1.4 Let $(X, \mathcal{F}_X)$ and $(Y, \mathcal{F}_Y)$ be two measurable spaces. A function $f : X \to Y$ is said to be *measurable* if for each measurable set $B \in \mathcal{F}_Y$, the inverse image $f^{-1}(B) \in \mathcal{F}_X$.

Proposition 1.5 Let $(X, \mathcal{F}_X)$, $(Y, \mathcal{F}_Y)$, and $(Z, \mathcal{F}_Z)$ be measurable spaces, and let $f : X \to Y$ and $g : Y \to Z$ be measurable functions. Then $g \circ f : X \to Z$ is measurable.

Definition 1.6 A *probability space* is a measure space $(\Omega, \mathcal{F}, P)$ with unit total measure (that is, $P(\Omega) = 1$). It is used to model a real world stochastic process.

Ω is the set of all possible outcomes for a single execution of the process, and is known as the *sample space*.

Sets $A \in \mathcal{F}$ are called *events*.

P is known as the *probability measure*. Note that $P(\Omega) = 1$, and P is a measure and hence is nonnegative and countably additive. Thus, it follows that for all events A, $P(A) \in [0, 1]$.

1.2 General Theory

Definition 1.7 Let $(\Omega, \mathcal{F}, P)$ be a probability space. A *random variable* is a measurable function $X : \Omega \to E$ for some measurable space $(E, \mathcal{F}_E)$. The probability that X takes on a value in a measurable set $S \in \mathcal{F}_E$ is denoted

$$P(X \in S)$$

and is given by

$$P(X \in S) = P\left(\{\omega \in \Omega \mid X(\omega) \in S\}\right).$$

If S is the singleton $S = \{s\}$, this is also sometimes written as

$$P(X = s).$$

Definition 1.8 Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A \in \mathcal{F}$ be an event. The *indicator function* of A is a function $\mathbb{1}_A : \Omega \to \{0, 1\}$ defined by

$$\mathbb{1}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

Proposition 1.9 Let $A, B \subseteq X$ be sets. The indicator function has the following properties:

- $\mathbb{1}_{A \cap B} = \min\{\mathbb{1}_A, \mathbb{1}_B\} = \mathbb{1}_A \cdot \mathbb{1}_B$
- $\mathbb{1}_{A \cup B} = \max\{\mathbb{1}_A, \mathbb{1}_B\} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_A \cdot \mathbb{1}_B$
- $\mathbb{1}_{A'} = 1 - \mathbb{1}_A$

Definition 1.10 Let $X : \Omega \to \mathbb{R}$ be a real-valued random variable. The *cumulative distribution function*, or *distribution function* of X, often abbreviated to *cdf* is the function

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

The probability that X is contained in the interval $(a, b]$ is therefore

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a).$$

Definition 1.11 Let A be an event such that $P(A) = 1$. In this case A is said to happen *almost surely*.

Definition 1.12 Let A be an event such that A' happens almost surely (or, equivalently, $P(A) = 0$). In this case A is said to happen *almost never*.

Proposition 1.13 Let $(\Omega, \mathcal{F}, P)$ be a probability space and let A be an event. Then

$$P(A') = 1 - P(A).$$

Proposition 1.14 Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then

$$P(\emptyset) = 0.$$

Proposition 1.15 Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let A and B be events. If $A \subseteq B$, then $P(A) \leq P(B)$.

Proposition 1.16 Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let A, B, and C be events. Then

$$P(A \cup B) = P(A) + P(B) + P(A \cap B)$$

and

$$\begin{aligned} P(A \cup B \cup C) = P(A) + P(B) + P(C) \\ - P(A \cap B) - P(B \cap C) - P(A \cap C) \\ + P(A \cap B \cap C) \end{aligned}$$

1.3 Conditional Probability

Definition 1.17 Let A and B be events. The probability that A will happen given that B has already happened is called the *probability of A given B*, is denoted $P(A \mid B)$ and is given by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Definition 1.18 Two events A and B are said to be independent if

$$P(A \mid B) = P(A)$$

or, equivalently,

$$P(A \cap B) = P(A)P(B).$$

Theorem 1.19 Let A and B be independent events. Then

- A' and B are independent,
- A and B' are independent, and
- A' and B' are independent.

Theorem 1.20 – Bayes' Theorem Let $B_1, \cdots, B_n$ be a partition of the sample space Ω (that is, $B_1, \cdots, B_n$ are mutually exclusive and exhaustive), and let A be an event. Then

$$P(B_k \mid A) = \frac{PA \mid B_k)P(B_k)}{\sum_{i=1}^{m} P(A \mid B_i)P(B_i)}, \quad k = 1, 2, \ldots, n.$$

SECTION 2
COMBINATORICS

Theorem 2.1 Suppose we are to randomly select r people out of a population of n without replacement and such that order matters. This is referred to as a *permutation*. The number of ways to do this is denoted ${}_nP_r$ and is given by

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

.

Theorem 2.2 Suppose we are to randomly select r people out of a population of n without replacement and that order does not matter. The number of ways to do this is given by the binomial coefficient

$$ {}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!} $$

Theorem 2.3 – **Vandermonde's Identity** Let $r, m, n \in \mathbb{N}$. Then

$$ \binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k}. $$

More generally,

$$ \binom{n_1 + \cdots + n_p}{m} = \sum_{k_1+\cdots+k_p=m} \binom{n_1}{k_1}\binom{n_2}{k_2}\cdots\binom{n_p}{k_p}. $$

Theorem 2.4 – **Pascal's rule** Binomial coefficients can be calculated recursively by

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

This is used to generate *Pascal's triangle* where each vertex is the sum of the nearest two vertices above it, generating the binomial coefficients.

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & / & & \backslash & & & & \\
 & & & 1 & & & & 1 & & & \\
 & & / & & \backslash & & / & & \backslash & & \\
 & & 1 & & & 2 & & & 1 & & \\
 & / & & \backslash & / & & \backslash & / & & \backslash & \\
 & 1 & & 3 & & & 3 & & & 1 & \\
/ & & \backslash & / & \backslash & & / & \backslash & / & & \backslash \\
1 & & 4 & & & 6 & & & 4 & & 1 \\
\dots & & \dots & & \dots & & \dots & & \dots & & \dots
\end{array}$$

Definition 2.5 Let $X_1, X_2, \dots, X_n$ be independent random variables which all share the same distribution. In this case the random variables $X_1, \dots, X_n$ are said to be *independent and identically distributed*, or *i.i.d.* for short.

SECTION 3

PROPERTIES OF RANDOM VARIABLES AND DISTRIBUTIONS

3.1 Expectation and Moments

Definition 3.1 Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$. Then the *expected value* or *expectation* of X is defined by the Lebesgue integral

$$\mathbb{E}[X] = \int_{\Omega} X(\omega) dP(\omega).$$

If X is a real valued random variable with cumulative distribution function F. The expectation of X can be written as

$$E[X] = \int\limits_{-\infty}^{\infty} x dF(x).$$

If

$$E[|X|] = \int\limits_{-\infty}^{\infty} |x| dF(x) = \infty$$

then the expectation of X is said not to exist. The expectation of X is sometimes denoted $\langle x \rangle$.

Definition 3.2 Let X be a real valued random variable with cumulative distribution function F. The n^{th} *moment* of X is the expectation of the random variable X^n,

$$\mu_n = \mathbb{E}[X^n] = \int_{-\infty}^{\infty} x^n dF(x).$$

Similarly, if

$$E[|X^n|] = \int_{-\infty}^{\infty} |x^n| dF(x) = \infty$$

then the n^{th} moment of X is said not to exist.

Note that this means the expectation of X is equal to the first moment of X.

Definition 3.3 Let X be a real valued random variable with cumulative distribution function F. The *mean* of X, denoted μ_X or simply μ, is defined to be the first moment, or expectation, of X.

$$\mu_X = \mathbb{E}[X] = \int_{-\infty}^{\infty} x dF(x)$$

Definition 3.4 Let X be a real valued random variable with mean μ and cumulative distribution function F. The n^{th} *central moment* of X is defined to as

$$E[(X-\mu)^n] = \int_{-\infty}^{\infty} (x-\mu)^n dF(x).$$

As above, if $E[|X-\mu|^n] = \infty$, the n^{th} central moment of X is said not to exist.

Definition 3.5 Let X be a real valued random variable with mean μ and cumulative distribution function F. The *variance* of X, denoted σ^2 is the second central moment of X. That is,

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 dF(x) = \mathbb{E}[(X-\mu)^2]$$

The value σ is known as the *standard deviation* of X.

Definition 3.6 Let X be a real valued random variable with mean μ, variance σ^2, and cumulative distribution function F. The n^{th} *standardized moment* is the n^{th} central moment divided by σ^n. That is, it is given by

$$\frac{E[(X-\mu)^n]}{\sigma^n}.$$

Proposition 3.7 Let X and Y be real valued random variables and let $a, b, c \in \mathbb{R}$. Then

$$E[aX + bY + c] = aE[X] + bE[Y] + c.$$

Proposition 3.8 Let X be a real valued random variable with mean μ. Then the variance of X is given by

$$\sigma^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mu^2.$$

Definition 3.9 Let X be a real valued random variable. The *skewness*, *skewness coefficient*, or *Pearson moment* of X is defined as the third standardized moment of X,

$$\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mathbb{E}[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}.$$

The skewness is a measure of the asymmetry of a distribution about its mean.

Definition 3.10 Let X be a real valued random variable. The *kurtosis* of X is defined as the fourth standardized moment of X,

$$\text{Kurt}[X] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right].$$

The kurtosis of X is a measure of how heavy-tailed or light-tailed its distribution is (that is, how slowly or quickly its probability drops off as $x \to \pm\infty$). It can also be thought of as the distributions propensity to produce outliers, and how extreme those outliers tend to be.

Definition 3.11 A normal distribution has a kurtosis of 3. It is common to compare other distributions to this result. Let X be a real valued random variable. The *excess kurtosis* of X is defined to be the difference

$$\text{Kurt}[X] - 3.$$

This breaks distributions into three regimes:

- A distribution is said to be *platykurtic* if it has negative excess kurtosis. Distributions in this regime will be more light tailed and produce fewer outliers than a normal distribution. An example of this would be the uniform distribution, which does not produce outliers. Distributions in this regime are sometimes called *sub-Gaussian*.
- A distribution is said to be *mesokurtic* if it has no excess kurtosis. An example of this would be the binomial distribution with $p = \frac{1}{2} \pm \frac{1}{\sqrt{12}}$.
- A distribution is said to be *leptokurtic* if it has positive excess kurtosis. Distributions in this regime will be heavier tailed and produce more outliers than a normal distribution. An example of this would be a Poisson distribution. Distributions in this regime are sometimes called *super-Gaussian*.

3.2 Moment Generating Function

Definition 3.12 Let X be a real valued random variable. The *moment generating function* of X is defined as

$$M_X(t) = \mathbb{E}[e^{tX}], \quad t \in \mathbb{R}.$$

Note that by applying the Taylor expansion of e^{tX}, we have

$$\begin{aligned}
M_X(t) &= \mathbb{E}[e^{tX}] \\
&= \mathbb{E}\left[1 + tX + \frac{t^2X^2}{2} + \cdots + \frac{t^mX^m}{m!} + \cdots\right] \\
&= \mathbb{E}[1] + \mathbb{E}[X] + \mathbb{E}\left[\frac{t^2X^2}{2}\right] + \cdots \mathbb{E}\left[\frac{t^mX^m}{m!}\right] + \cdots \\
&= 1 + \mathbb{E}[X] + \frac{t^2}{2}\mathbb{E}[X^2] + \cdots + \frac{t^m}{m!}\mathbb{E}[X^m] + \cdots \\
&= \sum_{m=0}^{\infty} \frac{t^m}{m!}\mathbb{E}[X^m]
\end{aligned}$$

so that the n^{th} derivative of $M_X(t)$ at $t = 0$ gives the n^{th} moment of X.

$$\begin{aligned} M_X^{(n)}(0) &= \left(\sum_{m=0}^{\infty} \frac{t^{m-n}}{(m-n)!} \mathbb{E}[X^m] \right) \Big|_{t=0} \\ &= \left(\sum_{m=0}^{\infty} \frac{t^m}{m!} \mathbb{E}[X^{m+n}] \right) \Big|_{x=0} \\ &= \sum_{m=0}^{\infty} \frac{0^m}{m!} \mathbb{E}[X^{m+n}] \\ &= \mathbb{E}[X^n] \end{aligned}$$

Proposition 3.13 The moment generating function has the following properties.

- Let X and Y be any two real valued random variables. Then X and Y are identically distributed if and only if $M_X(t) = M_Y(t)$ for all $t \in \mathbb{R}$.

- Let $X_1, \dots, X_n$ be independent random variables and let $a_1, \dots, a_n, b$ be constants. Define a random variable $Y = a_1 X_1 + \cdots + a_n X_n + b$. Then

$$M_Y(t) = M_{a_1X_1+\cdots+a_nX_n+b}(t) = e^{bt} M_{X_1}(a_1 t) \cdots M_{X_n}(a_n t).$$

3.3 Characteristic Function

Definition 3.14 Let X be a real valued random variable. The *characteristic function* of X is defined as

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right], \;\; t \in \mathbb{R}.$$

Proposition 3.15 The characteristic function has the following properties.

- The characteristic function of a real valued random variable always exists.

- Let X_1 and X_2 be any two real valued random variables. Then X_1 and X_2 are identically distributed if and only if $\varphi_{X_1}(t) = \varphi_{X_2}(t)$.

- If a random variable X admits a probability density $f(x)$, then its characteristic function is the Fourier transform of f.

- If a random variable admits a moment generating function $M_X(t)$, then $M_X(t) = \varphi_X(-it)$.
- Let $X_1, \ldots, X_n$ be independent random variables and let $a_1, \ldots, a_n, b$ be constants. Define a random variable $Y = a_1X_1 + \cdots + a_nX_n + b$. Then

$$\varphi_Y(t) = \varphi_{a_1X_1+\cdots+a_nX_n+b}(t) = e^{itb}\varphi_{X_1}(a_1t)\cdots\varphi_{X_n}(a_nt).$$

Proposition 3.16 Let X be a real-valued random variable and let A be an event. Then $P(A)$ can be expressed as

$$P(A) = P(X \in A) = \mathbb{E}\left[\mathbb{1}_A\right].$$

In particular, the cumulative distribution function $F_X(x)$ can be expressed as

$$F_X(x) = P(X \leq x) = \mathbb{E}\left[\mathbb{1}_{\{X \leq x\}}\right].$$

Definition 3.17 Let X_1 and X_2 be independent copies of a random variable X. The distribution of X is said to be a *stable distribution* if for any constants $a, b > 0$, there exists some $c > 0$ and $d \in \mathbb{R}$ such that $aX_1 + bX_2$ shares the same distribution as $cX + d$.

For example, the normal distribution $\mathcal{N}(\mu, \sigma)$ is stable.

SECTION 4

DISCRETE DISTRIBUTIONS

Definition 4.1 Let $(\Omega, \mathcal{F}, P)$ be a probability space. A *discrete random variable* is a random variable $X : \Omega \to E$ such that $X(\Omega)$ is countable. A common case is a random variable $X : \Omega \to \mathbb{Z}$.

Definition 4.2 Let $X : \Omega \to E$ be a discrete random variable. The *probability mass function* of X is a function $p_X : X(\Omega) \subseteq E \to [0, 1]$ defined by

$$p_X(x_i) = P(X = x_i)$$

where P is the probability measure of the probability space $(\Omega, \mathcal{F}, P)$. When no confusion can occur, we drop the subscript and write $p_X(x)$ as simply write $p(x)$.

Remark 4.3 Let X be a real valued discrete random variable. We have the following identities:

- $\mu_X = \mathbb{E}[X] = \sum_x x \cdot p_X(x)$
- $\mathbb{E}[X \mid Y] = \sum_x x \cdot P(X = x \mid Y)$
- $\sigma^2 = \sum_x (x - \mu)^2 p_X(x) = \left(\sum_x x^2 p_X(x) \right) - \mu^2$
- $E[X^n] = \sum_x x^n p_X(x)$
- $M_X(t) = \sum_x e^{tx} p_X(x)$

4.1 Bernoulli Distribution

4.1.1 Interpretation

A Bernoulli trial is a stochastic process for which the outcome is one of two possible values (typically 0 or 1, true or false, success or fail, etc) with probabilities p and $q = 1 - p$ of getting each result respectively.

4.1.2 Properties

Parameters	$p \in [0, 1]$ $q = 1 - p$
Support	$k \in \{0, 1\}$
Probability Mass Function	$p(k) = \begin{cases} q = 1 - p, & k = 0 \\ p, & k = 1 \end{cases}$
Cumulative Distribution Function	$F(k) = \begin{cases} 0, & k < 0 \\ 1 - p, & 0 \leq k < 1 \\ 1, & k \geq 1 \end{cases}$
Mean	p
Median	$\begin{cases} 0, & p < \frac{1}{2} \\ \{0, 1\}, & p = \frac{1}{2} \\ 1, & p > \frac{1}{2} \end{cases}$
Mode	$\begin{cases} 0, & p < \frac{1}{2} \\ \{0, 1\}, & p = \frac{1}{2} \\ 1, & p > \frac{1}{2} \end{cases}$
Variance	$p(1 - p) = pq$
Skewness	$\frac{q - p}{\sqrt{pq}} = \frac{1 - 2p}{\sqrt{pq}}$
Excess Kurtosis	$\frac{1 - 6pq}{pq}$
Entropy	$-q \ln q - p \ln p$
Moment Generating Function	$M(t) = q + pe^t$
Characteristic Function	$\varphi(t) = q + pe^{it}$
Probability Generating Function	$G(z) = q + pz$
Fisher Information	$\frac{1}{pq}$

4.2 Binomial Distribution

4.2.1 Interpretation

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed Bernoulli trials. The binomial distribution with n trials and probability of success p represents the probability of getting a given number of successes among the n trials. Equivalently the binomial distribution has the same distribution as $X_1 + \cdots + X_n$.

4.2.2 Properties

Notation	$B(n, p)$
Parameters	$n \in \mathbb{N}$ $p \in [0, 1]$ $q = 1 - p$
Support	$k \in \{0, \ldots, n\}$
Probability Mass Function	$p(k) = \binom{n}{k} p^k q^{n-k}$
Cumulative Distribution Function	$F(k) = \sum_{i=0}^{k} \binom{n}{i} p^i q^{n-i}$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil + 1$
Variance	npq
Skewness	$\frac{q - p}{\sqrt{npq}}$
Excess Kurtosis	$\frac{1 - 6pq}{npq}$
Entropy	$\frac{1}{2} \log_2(2\pi enpq) + O\left(\frac{1}{n}\right)$
Moment Generating Function	$M(t) = (q + pe^t)^n$
Characteristic Function	$\varphi(t) = (q + pe^{it})^n$
Probability Generating Function	$G(z) = (q + pz)^n$
Fisher Information	$\frac{n}{pq}$

4.2.3 Sum of Binomials

Let $X \sim B(n, p)$ and $Y \sim B(m, p)$ be independent binomial random variables and define $Z = X + Y$. Then $Z \sim B(n + m, p)$.

4.2.4 Ratio of Binomials

Let $X \sim B(n, p_1)$ and $Y \sim B(m, p_2)$ be independent, and define $T = \frac{\frac{1}{n}X}{\frac{1}{m}Y} = \frac{mX}{nY}$. Then $\log(T)$ is approximately normally distributed with mean $\log(\frac{p_1}{p_2})$ and variance $\frac{\frac{1}{p_1}-1}{n} + \frac{\frac{1}{p_2}-1}{m}$ ([?]).

4.2.5 Conditional Binomials

Let $X \sim B(n, p)$ and let $Y|X \sim B(X, q)$. Then $Y \sim B(n, pq)$.

4.2.6 Normal Approximation

Let $X \sim B(n, p)$. In the limit as n become large, X can be approximated as a normal distribution $\mathcal{N}(np, np(1-p))$. This approximation works best when $n > 20$ and p is not near 0 or 1. Some common rules of thumb for deciding whether this approximation is appropriate are

- $n > 5$, and the skewness is less than $\frac{1}{3}$ in absolute value. That is,

$$\frac{|1-2p|}{\sqrt{np(1-p)}} = \frac{1}{\sqrt{n}}\left|\sqrt{\frac{1-p}{p}} - \sqrt{\frac{p}{1-p}}\right| < \frac{1}{3}.$$

- $\mu \pm 3\sigma = np \pm 3\sqrt{np(1-p)} \in (0, n)$ or, equivalently,

$$n > 9\left(\frac{1-p}{p}\right) \quad \text{and} \quad n > 9\left(\frac{p}{1-p}\right)$$

  which together imply the above criterion.

- Both np and $n(1-p)$ are greater than some chosen constant. A common choice is 5, however choosing 9 implies the above two criteria.

4.2.7 Poisson Approximation

The binomial distribution $B(n, p)$ converges toward the Poisson distribution with parameter $\lambda = np$ as $n \to \infty$ while the product np remains fixed (or $p \to 0$). Two common rules of thumb for deciding whether this approximation is appropriate are

- $n \geq 20$ and $p \leq 0.05$, and
- $n \geq 100$ and $np \leq 10$.

4.3 Geometric Distribution

4.3.1 Interpretation

The geometric distribution models the number of failures of successive independent and identically distributed Bernoulli trials with probability p that are obtained before obtaining one success.

4.3.2 Properties

Notation	$Geo(p)$
Parameters	$p \in [0, 1]$
Support	$k \in \mathbb{N}$
Probability Mass Function	$p(k) = (1-p)^k p$
Cumulative Distribution Function	$F(k) = 1 - (1-p)^{k+1}$
Mean	$\frac{1-p}{p}$
Median	$\left\lceil \frac{-1}{\log_2(1-p)} \right\rceil - 1$
Mode	0
Variance	$\frac{1-p}{p^2}$
Skewness	$\frac{2-p}{\sqrt{1-p}}$
Excess Kurtosis	$6 + \frac{p^2}{1-p}$
Entropy	$\frac{-(1-p)\log_2(1-p) - p\log_2 p}{p}$
Moment Generating Function	$M(t) = \frac{p}{1-(1-p)e^t}$
Characteristic Function	$\varphi(t) = \frac{p}{1-(1-p)e^{it}}$
Probability Generating Function	$G(z) = \frac{p}{1-z(1-p)}$

4.3.3 Memorylessness

The geometric distribution is memoryless. That is, if X is a geometric random variable and $m, n \in \mathbb{N}$ are any positive integers, then

$$P(X > m + n \mid X > n) = P(X > m).$$

4.3.4 Sum of Geometric Random Variables

Let $X_1, \ldots, X_r$ be independent and identically distributed random variables with distribution $Geo(p)$, and define a new random variable $Y = \sum_{i=1}^{r} X_i$. Then Y follows a negative binomial distribution with parameters r and p. In particular, this means the geometric distribution is the negative binomial distribution with $r = 1$.

4.3.5 Minimum of Geometrics Random Variables

Let $X_1, \ldots, X_n$ be independent geometrically distributed random variables with (possibly distinct) success parameters p_i, and define a new random variable $Y = \min_{i \in 1, \cdots, n} Y_i$. Then W is also geometrically distributed with parameter $p = 1 - \prod_i (1 - p_i)$.

4.4 Negative Binomial Distribution

4.4.1 Interpretation

The negative binomial distribution $NB(r, p)$ models the number of failures in a sequence of independent and identically distributed Bernoulli trials with probability of success p before a specified number of successes r occur.

4.4.2 Properties

Notation	$NB(r, p)$
Parameters	$r \in \mathbb{N}_+$ $p \in [0, 1]$
Support	$k \in \mathbb{N}$
Probability Mass Function	$p(k) = \binom{k+r-1}{k}(1-p)^k p^r$
Cumulative Distribution Function	$F(k) = \sum_{i=0}^{k} \binom{r+i-1}{i} p^r q^i$
Mean	$\frac{pr}{1-p}$
Mode	$\begin{cases} \lfloor \frac{p(r-1)}{1-p} \rfloor, & r > 1 \\ 0, & r \leq 1 \end{cases}$
Variance	$\frac{pr}{(1-p)^2}$
Skewness	$\frac{1+p}{\sqrt{pr}}$
Excess Kurtosis	$\frac{6}{r} + \frac{(1-p)^2}{pr}$
Moment Generating Function	$M(t) = \left(\frac{1-p}{1-pe^t}\right)^r$, $t < -\ln p$
Characteristic Function	$\varphi(t) = \left(\frac{1-p}{1-pe^{it}}\right)^r$, $t \in \mathbb{R}$
Probability Generating Function	$G(z) = \left(\frac{1-p}{1-pz}\right)^r$, $\lvert z \rvert < \frac{1}{p}$
Fisher Information	$\frac{r}{(1-p)^2 p}$
Method of Moments	$r = \frac{\mathbb{E}[X]^2}{\mathrm{Var}[X] - \mathbb{E}[X]}$ $p = 1 - \frac{\mathbb{E}[X]}{\mathrm{Var}[X]}$

4.5 Discrete Uniform Distribution

4.5.1 Interpretation

A discrete uniform random variable models a process where a finite number of values between two numbers a and b are equally likely outcomes, such as rolling a fair six-sided die.

4.5.2 Properties

Notation	$\mathcal{U}(a,b)$ or $\mathrm{unif}(a,b)$
Parameters	$a, b \in \mathbb{Z}$ with $a \leq b$ $n = b - a + 1$
Support	$k \in \{a, a+1, \ldots, b-1, b\}$
Probability Mass Function	$p(k) = \frac{1}{n}$
Cumulative Distribution Function	$F(k) = \frac{k-a+1}{n}$
Mean	$\frac{a+b}{2}$
Median	$\frac{a+b}{2}$
Median	N/A
Variance	$\frac{(b-a+1)^2-1}{12}$
Skewness	0
Excess Kurtosis	$-\frac{6(n^2+1)}{5(n^2-1)}$
Entropy	$\ln(n)$
Moment Generating Function	$M(t) = \frac{e^{at} - e^{(b+1)t}}{b(1-e^t)}$
Characteristic Function	$\varphi(t) = \frac{e^{iat} - e^{i(b+1)t}}{b(1-e^{it})}$
Probability Generating Function	$G(z) = \frac{z^a - z^{b+1}}{n(1-z)}$

4.6 Hypergeometric Distribution

4.6.1 Interpretation

The hypergeometric distribution models the number, k, of successes in n draws without replacement from a finite population of size N containing K objects of the desired type. An example of this would be drawing $n = 5$ times at random from a jar containing $N = 10$ marbles, $K = 6$ of which are red and $N - K = 4$ of which are green, and counting how many, k, are red. In contrast, the binomial distribution models the number of successes with replacement (i.e. when the drawn marble is put back into the jar between draws).

4.6.2 Properties

Parameters	$N \in \mathbb{N}$ $K \in \{0, 1, \ldots, N\}$ $n \in \{0, 1, \ldots, N\}$
Support	$\max(0, n + K - N) \leq k \leq \min(n, K)$
Probability Mass Function	$p(k) = \frac{\binom{K}{k}\binom{N-n}{n-k}}{\binom{N}{n}}$
Mean	$\frac{nK}{N}$
Median	$\left\lceil \frac{(n+1)(K+1)}{(N+2)} \right\rceil - 1,$ or $\left\lfloor \frac{(n+1)(K+1)}{(N+2)} \right\rfloor$
Variance	$n\frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$
Skewness	$\frac{(N-2K)(N-2n)\sqrt{N-1}}{(N-2)\sqrt{nK(N-K)(N-n)}}$

4.6.3 Symmetries

The hypergeometric distribution admits the following symmetries:

- swapping the role of red and green marbles

$$f(k; N, K, n) = f(n - k; N, N - K, n)$$

- swapping the role of drawn and not drawn marbles

$$f(k; N, K, n) = f(K - k, N, K, N - n)$$

- swapping the roles of green marbles and drawn marbles

$$f(k; N, K, n) = f(k; N, n, K)$$

4.6.4 Binomial Approximation of Hypergeometric Distributions

Let X be a hypergeometric random variable with parameters N, K, and n, let $p = \frac{K}{N}$, and let $Y \sim B(n, p)$. If $N \geq K \gg n$ and p is not close to 0 or 1, then X and Y have approximately the same distribution so that

$$P(X \leq k) \simeq P(Y \leq k).$$

4.6.5 Normal Approximation of Hypergeometric Distributions

If n is large and $N, K \gg n$, and $p = \frac{K}{N}$ is not close to 0 or 1, then

$$P(X \leq k) \approx \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right)$$

where Φ is the standard normal distribution function.

4.7 Poisson Distribution

4.7.1 Interpretation

The Poisson distribution models the number of of events occurring within a fixed interval given that these events occur with a known constant rate, λ, on average and independently of the time that the last event occurred. For example, one might use the Poisson distribution to describe the number of defects produced in 30 yards of fabric given that on average 1 defect is produced per 5 yards. In this case, $\lambda = 1 \cdot \frac{30}{5} = 6$.

4.7.2 Properties

Notation	$\mathrm{Pois}(\lambda)$
Parameters	$n \in \mathbb{N}$ $\lambda \in (0, \infty)$
Support	$k \in \mathbb{N}$
Probability Mass Function	$p(k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$
Cumulative Distribution Function	$F(k) = e^{-\lambda} \sum_{i=0}^{k} \dfrac{\lambda^i}{i!}$
Mean	λ
Median	$\approx \left\lfloor \lambda + \frac{1}{3} - \frac{0.02}{\lambda} \right\rfloor$
Mode	$\lfloor \lambda \rfloor$
Variance	λ
Skewness	$\dfrac{1}{\sqrt{\lambda}}$
Excess Kurtosis	$\dfrac{1}{\lambda}$
Entropy	$\lambda(1 - \ln \lambda) + e^{-\lambda} \sum_{k=0}^{\infty} \dfrac{\lambda^k \ln(k!)}{k!}$
Moment Generating Function	$M(t) = e^{\lambda(e^t - 1)} = \exp[\lambda(e^t - 1)]$
Characteristic Function	$\varphi(t) = e^{\lambda(e^{it} - 1)} = \exp[\lambda(e^{it} - 1)]$
Probability Generating Function	$G(z) = e^{\lambda(z-1)} = \exp[\lambda(z - 1)]$
Fisher Information	$\dfrac{1}{\lambda}$

4.7.3 Sum of Poisson Random Variables

Let $X_i \sim \text{Pois}(\lambda_i)$ for $i = 1, \ldots, n$ be independent and define a new random variable $Y = \sum_i X_i$. Then

$$Y \sim \text{Pois}\left(\sum_i \lambda_i\right).$$

Theorem 4.4 – Raikov's Theorem Let $Z \sim \text{Pois}(\lambda_Z)$ be a random variable and suppose that there are independent random variables X and Y such that $Z = X + Y$. Then the distribution of X and Y are both a shifted Poisson distribution with parameters λ_X and λ_Y, respectively. Moreover, $\lambda_X + \lambda_Y = \lambda_Z$.

4.7.4 Conditional Poisson Distributions

Let $X \sim \text{Pois}(\lambda_X)$ and $Y \sim \text{Pois}(\lambda_Y)$ be independent random variables. Define a random variable $Z = X \mid X + Y$. Then Z follows a binomial distribution. Specifically, if $X + Y = k$, then $Z \sim B\left(k, \dfrac{\lambda_X}{\lambda_X + \lambda_Y}\right)$.

Now, let $X \sim \text{Pois}(\lambda_X)$ and suppose $Y \mid (X = k) \sim B(k, p)$. Then Y follows a Poisson distribution $Y \sim \text{Pois}(\lambda p)$.

4.7.5 Normal Approximation

For large values of *lambda* (starting around 1000 or so), the Poisson distribution with parameter λ is very closely approximated by a normal distribution with mean λ and variance λ. However, for $\lambda \geq 10$ or so, the Poisson distribution can still be approximated by a normal distribution if a continuity correction is performed.

4.7.6 Variance-Stabilizing Transformation

Let $X \sim \text{Pois}(\lambda)$. Then

$$Y = 2\sqrt{X} \approx \mathcal{N}(2\sqrt{\lambda}, 1)$$

and

$$Z = \sqrt{X} \approx \mathcal{N}\left(\sqrt{\lambda}, \frac{1}{4}\right).$$

4.8 Kronecker Delta

Definition 4.5 The *Kronecker delta* is a function of two variables, typically two real numbers. It can be thought of as the indicator function for the event that the two variables are equal. That is

$$\delta_{ij} = \mathbb{1}_{\{i=j\}} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

The Kronecker delta is the discrete analog of the Dirac delta.

4.8.1 Properties

- $\sum_i a_i \delta_{ij} = a_j$
- $\sum_j \delta_{ij} a_j = a_i$
- $\sum_k \delta_{ik} \delta_{kj} = \delta_{ij}$

SECTION 5

CONTINUOUS DISTRIBUTIONS

Definition 5.1 Let $(\Omega, \mathcal{F}, P)$ be a probability space. A *continuous random variable* is a random variable $X : \Omega \to E$ such that $X(\Omega)$ is uncountably infinite.

Definition 5.2 Let $X : \Omega \to E$ be a continuous random variable. The *probability density function* of X is a function $f_X : X(\Omega) \subseteq E \to \mathbb{R}_+$ defined by

Remark 5.3 Let X be a real valued continuous random variable. We have the following identities:

- $\mu_X = \mathbb{E}[X] = \int\limits_{-\infty}^{\infty} x \cdot f_X(x)dx$
- $\sigma^2 = \int\limits_{-\infty}^{\infty} (x - \mu)^2 f_X(x)dx = \left(\int_{-\infty}^{\infty} x^2 f_X(x)dx\right) - \mu^2$
- $E[X^n] = \int\limits_{-\infty}^{\infty} x^n f_X(x)dx$
- $M_X(t) = \int\limits_{-\infty}^{\infty} e^{tx} f_X(x)dx$

5.1 Continuous Uniform Distribution

5.1.1 Interpretation

The continuous uniform distribution models a process which chooses a random real number in a designated interval $x \in [a, b]$, with no areas being more likely than others.

5.1.2 Properties

Notation	$\mathcal{U}(a, b)$
Parameters	$a < b \in \mathbb{R}$
Support	$x \in [a, b]$
Probability Density Function	$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a, b] \\ 0, & \text{else} \end{cases}$
Cumulative Distribution Function	$F(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \in [a, b] \\ 1, & x > b \end{cases}$
Mean	$\dfrac{a+b}{2}$
Median	$\dfrac{a+b}{2}$
Mode	N/A
Variance	$\frac{1}{12}(b-a)^2$
Skewness	0
Excess Kurtosis	$\dfrac{-6}{5}$
Entropy	$\ln(b-a)$
Moment Generating Function	$M(t) = \begin{cases} \dfrac{e^{tb} - e^{ta}}{t(b-a)}, & t \neq 0 \\ 1, & t = 0 \end{cases}$
Characteristic Function	$\varphi(t) = \begin{cases} \dfrac{e^{itb} - e^{ita}}{it(b-a)}, & t \neq 0 \\ 1, & t = 0 \end{cases}$

Definition 5.4 In this case when $a = 0$ and $b = 1$, we have the distribution $\mathcal{U}(0,1)$. This is called the *standard uniform distribution*.

The standard uniform distribution can be used to generate random numbers from any distribution. Let u be a random number generated from $\mathcal{U}(0,1)$. Then $x = F^{-1}(u)$ generates a random number x from a continuous random variable with cumulative distribution function F.

5.1.3 Powers of the Standard Uniform

Let $X \sim \mathcal{U}(0,1)$ and define $Y = X^n$. Then Y has a beta distribution $\text{Beta}\left(\frac{1}{n}, 1\right)$.

5.1.4 Sum of Uniform Distributions

Let $X_1, \ldots, X_n$ be independent and identically distributed $\mathcal{U}(0,1)$ random variables and define $Y = X_1 + \cdots + X_n$. Then Y has an Irwin-Hall distribution.

5.2 Exponential Distribution

5.2.1 Interpretation

The exponential distribution models the time between events in a Poisson process. If $X \sim \text{Pois}(\lambda)$, then the time until the first event and the time between successive events have the exponential distribution with parameter λ.

5.2.2 Properties

Notation	$\text{Exp}(\lambda)$
Parameters	$\lambda > 0$
Support	$x \in [0, \infty)$
Probability Density Function	$f(x) = \lambda e^{-\lambda x}$
Cumulative Distribution Function	$F(x) = 1 - e^{-\lambda x}$
$100Q^{th}$ Quantile	$-\frac{1-Q}{\lambda}$
Mean	$\frac{1}{\lambda}$
Median	$\frac{\ln 2}{\lambda}$
Mode	0
Variance	$\frac{1}{\lambda^2}$
Skewness	2
Excess Kurtosis	6
Entropy	$1 - \ln \lambda$
Moment Generating Function	$M(t) = \frac{\lambda}{\lambda - t}$, for $t < \lambda$
Characteristic Function	$\varphi(t) = \frac{i\lambda}{i\lambda - t}$

5.2.3 Memorylessness

If $T \sim \text{Exp}(\lambda)$ models the time for a Poisson event to occur, then the distribution of the waiting time until the next event is independent of the time already spent waiting for the event. That is, for all $s, t \geq 0$,

$$P(T > s + t \mid T > s) = P(T > t).$$

5.2.4 Minimum of Exponential Random Variables

Let $X_1, \dots, X_n$ be independent exponentially distributed random variables with parameters $\lambda_1, \dots, \lambda_n$. Let $Y = \min\{X_1, \dots, X_n\}$. Then Y is also exponentially distributed with parameter

$$\lambda = \lambda_1 + \cdots + \lambda_n.$$

5.2.5 Sum of Exponentials

Let X and Y be two independent exponentially distributed random variables with parameters λ_X and λ_Y respectively, and let $Z = X + Y$. Then the probability density function of Z is given by

$$f_Z(z) = \begin{cases} \dfrac{\lambda_X \lambda_Y}{\lambda_Y - \lambda_X} \left(e^{-\lambda_X z} - e^{-\lambda_Y z} \right), & \lambda_1 \neq \lambda_2 \\ \lambda^2 z e^{-\lambda z}, & \lambda_X = \lambda_Y = \lambda \end{cases}$$

Moreover, if $X_1, \dots, X_n$ be i.i.d. exponentially distributed random variables with rate parameter λ and define $Y = X_1 + \cdots + X_n$. Then

$$Y \sim \text{Gamma}(n, \lambda).$$

5.2.6 Relation to the Geometric Distribution

The exponential distribution is the continuous analogue of the geometric distribution. Let X be an exponentially distributed random variable with parameter λ and define $Y = \lfloor X \rfloor$. Then Y is geometrically distributed with parameter $p = 1 - e^{-\lambda}$.

5.3 Gamma Distribution

5.3.1 Interpretation

Among other things, the gamma distribution with parameters α and β models the waiting time until the α^{th} occurrence of a Poisson event in a Poisson process with parameter $\lambda = \beta$ [1].

5.3.2 Properties

Notation	$\text{Gamma}(\alpha, \beta)$
Parameters	$\alpha, \beta \in (0, \infty)$
Support	$x \in (0, \infty)$
Probability Density Function[2]	$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
Cumulative Distribution Function[3]	$F(x) = \frac{1}{\Gamma(\alpha)} \gamma(\alpha, \beta x)$
Mean	$\frac{\alpha}{\beta}$
Mode	$\frac{\alpha - 1}{\beta}$ for $\alpha \geq 1$
Variance	$\frac{\alpha}{\beta^2}$
Skewness	$\frac{2}{\sqrt{\alpha}}$
Excess Kurtosis	$\frac{6}{\alpha}$
Entropy[4]	$\alpha - \ln \beta + \ln \Gamma(\alpha) + (1 - \alpha)\psi(\alpha)$
Moment Generating Function	$M(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha}$, for $t < \beta$
Characteristic Function	$\varphi(t) = \left(1 - \frac{it}{\beta}\right)^{-\alpha}$

[1] It is also common to replace β with a parameter $\theta = \frac{1}{\beta}$.

[2] Here, $\Gamma(x)$ is the gamma function defined by $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$. Note that when $x \in \mathbb{N}$, $\Gamma(x) = x!$.

[3] Here, $\gamma(s, x)$ is the lower incomplete gamma function defined by $\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt$.

[4] Here, $\psi(\alpha)$ is the digamma function defined by $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$.

5.3.3 Sum of Gamma Distributions

let $X_i \sim \text{Gamma}(\alpha_i, \beta)$ be independent random variables for $i = 1, \ldots, n$ so that all distributions share the rate β with possibly varying α_i. Let $Y = X_1 + \cdots + X_n$ and define $\alpha = \alpha_1 + \cdots + \alpha_n$.. Then

$$Y \sim \text{Gamma}\left(k, \theta\right).$$

5.3.4 Scaling Gamma Distributions

Let $X \sim \text{Gamma}(\alpha, \beta)$, let $c > 0$, and define $Y = cX$. Then

$$Y \sim \text{Gamma}\left(\alpha, \frac{\beta}{c}\right).$$

5.3.5 Ratio of Gamma Distributions

Let $X_1 \sim \text{Gamma}(\alpha_1, \beta_1)$ and $X_2 \sim \text{Gamma}(\alpha_2, \beta_2)$ be independent. Then

$$\frac{\alpha_2 \beta_1 X_1}{\alpha_1 \beta_2 X_2} \sim F(2\alpha_1, 2\alpha_2).$$

5.3.6 Relation to Other Distributions

Let $X \sim \text{Gamma}(1, \lambda)$, then $X \sim \text{Exp}(\lambda)$

Let $X \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{1}{2}\right)$, then $X \sim \chi^2(\nu)$, the chi-square distribution with ν degrees of freedom.

Let X follow a Maxwell-Boltzmann distribution with parameter a. Then $X^2 \sim \text{Gamma}\left(\frac{3}{2}, 2a^2\right)$.

5.4 Chi-Square Distribution

5.4.1 Interpretation

The chi-square distribution with k degrees of freedom models the sum of squares of k independent standard normal random variables.

The chi-square distribution is a space case of the gamma distribution with parameters $\alpha = \frac{k}{2}$ and $\beta = \frac{1}{2}$.

5.4.2 Properties

Notation	$\chi^2(k)$ or χ^2_k
Parameters	$k \in \mathbb{N}_+$
Support	$x \in \begin{cases} (0,\infty), & k=1 \\ [0,\infty), & \text{else} \end{cases}$
Probability Density Function	$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$
Cumulative Distribution Function	$F(x) = \frac{1}{\Gamma(k/2)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
Mean	k
Median	$\approx k\left(1 - \frac{2}{9k}\right)^3$
Mode	$\max(k-2, 0)$
Variance	$2k$
Skewness	$\sqrt{\frac{8}{k}}$
Excess Kurtosis	$\frac{12}{k}$
Entropy	$\frac{k}{2} + \ln\left(2\Gamma\left(\frac{k}{2}\right)\right) + \left(1 - \frac{k}{2}\right)\psi\left(\frac{k}{2}\right)$
Moment Generating Function	$M(t) = (1-2t)^{-k/2}$, for $t < \frac{1}{2}$
Characteristic Function	$\varphi(t) = (1-2it)^{-k/2}$
Probability Generating Function	$G(z) = (1 - 2\ln z)^{-k/2}$ for $t \in (0, \sqrt{e})$

5.4.3 Sum of Chi-Squares

Let $X_i \sim \chi^2(k_i)$ be independent for $i \in 1, \ldots, n$ and define $Y = X_1 + \cdots + X_n$. Then

$$Y \sim \chi^2(k_1 + \cdots + k_n).$$

5.4.4 Sample Mean of Chi-Squares

Let $X_1, \ldots, X_n \sim \chi^2(k)$ be independent and identically distributed. Then

$$\overline{X} = \frac{1}{n} \sum_i X_i \sim \text{Gamma}\left(\alpha = \frac{nk}{2}, \beta = \frac{n}{2}\right).$$

5.5 Chi Distribution

5.5.1 Interpretation

The chi distribution with k degrees of freedom models the Euclidean norm of a vector of k independent standard normal random variables.

$$Y = \sqrt{\sum_{i=1}^{k} Z_i^2}$$

Thus, the square of a chi distribution with k degrees of freedom is a chi-square distribution with k degrees of freedom, so that $Y^2 \sim \chi^2(k)$. Then divided by $\sqrt{k-1}$, Y gives the unbiased estimate of the standard deviation of k samples taken from a standard normal population.

5.5.2 Properties

Notation	$\chi(k)$
Parameters	$k \in \mathbb{N}_+$ degrees of freedom
Support	$x \in [0, \infty)$
Probability Density Function	$f(x) = \dfrac{2^{1-k/2}}{\Gamma(k/2)} x^{k-1} e^{-x^2/2}$
Cumulative Distribution Function[5]	$P\left(\dfrac{k}{2}, \dfrac{x^2}{2}\right)$
Mean	$\mu = \dfrac{\sqrt{2}\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}$
Median	$\approx \sqrt{k\left(1 - \frac{2}{9k}\right)^3}$
Mode	$\sqrt{k-1}$
Variance	$k - \mu^2$
Skewness	$\gamma_1 = \dfrac{\mu}{\sigma^3}\left(1 - 2\sigma^2\right)$
Excess Kurtosis	$\dfrac{2}{\sigma^2}\left(1 - \mu\sigma\gamma_1 - \sigma^2\right)$

5.5.3 Absolute Standard Normal Distribution

Let $Z \sim \mathcal{N}(0,1)$. Then $|Z| \sim \chi(1)$.

[5] Here $P(x, y)$ denotes the regularized gamma function.

5.6 Normal Distribution

5.6.1 Interpretation

When there is a large number of observations, many variables measured in common situations will exhibit a bell curve. The canonical example of this is scholastic aptitude test and other test scores. In addition, under many circumstances, due to the central limit theorem, many distributions will converge to a normal distribution when averaged over a large number of samples.

5.6.2 Properties

Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$, the mean $\sigma^2 > 0$, the variance
Support	$x \in \mathbb{R}$
Probability Density Function	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Cumulative Distribution Function	$F(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]$
$100Q^{th}$ Quantile	$\mu + \sqrt{2}\sigma \text{erf}^{-1}(2Q - 1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2
Mean Absolute Deviation	$\sqrt{2\pi}\sigma$
Skewness	0
Excess Kurtosis	0
Entropy	$\frac{1}{2}\ln(2\pi e \sigma^2)$
Moment Generating Function	$M(t) = e^{\mu t + \sigma^2 t^2/2}$
Characteristic Function	$\varphi(t) = e^{i\mu t - \sigma^2 t^2/2}$
Fisher information	$\mathcal{I}(\mu, \sigma) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{bmatrix}$ $\mathcal{I}(\mu, \sigma^2) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}$

5.6.3 Standard Normal Distribution

The simplest and most commonly used case of the normal distribution is *standard normal distribution*, $\mathcal{N}(0,1)$. In this case, we deviate from the usual notation for the probability density and cumulative distribution functions, $f(x)$ and $F(x)$ respectively. Instead we denote the probability density function and cumulative density function of the standard normal distribution by

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

and

$$\Phi(z) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right]$$

respectively.

By convention, random variables with a standard normal distribution are typically denoted by Z. Note that for a general normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$f_X(x \mid \mu, \sigma^2) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right).$$

Moreover,

$$Z = \frac{X-\mu}{\sigma}$$

and so regardless of the values of μ and σ we need only ever know $\Phi(z)$ as

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) \\ &= P\left(Z \le \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right). \end{aligned}$$

5.6.4 Linear Transformations

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $a, b \in \mathbb{R}$. Define $Y = aX + b$. Then

$$Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

5.6.5 Sum of Normal Random Variables

Let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent, and define $Y = X_1 + X_2$. Then

$$Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Theorem 5.5 – Cramer's Theorem Let X_1 and X_2 be any two independent random variables and define $Y = X_1 + X_2$. Then Y follows a normal distribution if and only if X_1 and X_2 both follow normal distributions.

Theorem 5.6 – Bernstein's Theorem Let $X_1, \ldots, X_n$ be independent random variables, and define $Y_1 = \sum a_k X_k$ and $Y_2 = \sum b_k X_k$ any linear combination of the X_k. Then Y_1 and Y_2 are independent if and only if

- X_k is normally distributed for all k, and
- $\sum a_k b_k \sigma_k^2 = 0$ where σ_k^2 is the variance of X_k.

5.6.6 Central Limit Theorem

Theorem 5.7 – Central Limit Theorem Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with an arbitrary distribution, mean μ and variance σ^2. Define their average

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

and define

$$Z_n = \sqrt{n}\left(\overline{X}_n - \mu\right).$$

Then Z_n approximates a normal distribution with mean 0 and variance σ^2, and the approximation gets better as n increases. Indeed, in the limiting case,

$$Z = \lim_{n \to \infty} Z_n \sim \mathcal{N}(0, \sigma^2)$$

exactly follows a normal distribution with mean 0 and variance σ^2.

Corollary 5.8 As a result of Theorem **??**, any probability distribution which arises as the sum of some number of independent and identically distributed random variables can be approximated by a normal distribution for a large enough number of summands. For example, the binomial distribution $B(n, p)$, being the sum of n Bernoulli random variables, can be approximated by $\mathcal{N}(np, np(1-p))$ for large n and p not too close to 0 or 1. Similarly, the chi-square distribution $\chi^2(k)$, being the sum of squares of k standard normal random variables can be approximated by $\mathcal{N}(k, 2k)$ for large k.

5.7 Student's t-Distribution

5.7.1 Interpretation

The student's $t-$distribution, or simply the $t-$distribution, is used when estimating the mean of a normally distributed population when the sample size is small and the population standard deviation is unknown. The $t-$distribution with ν degrees of freedom models the distribution of the sample mean of $\nu + 1$ independent and identically distributed normal random variables relative to the true population mean (after multiplying by $\sqrt{n}$. It can also be used to assess the statistical significant of the difference between two sample means.

5.7.2 Properties

Parameters	$n \in \mathbb{N}$ degrees of freedom
Support	$x \in \mathbb{R}$
Probability Density Function	$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
Mean	$\begin{cases} 0, & \nu > 1 \\ \text{undefined}, & \text{else} \end{cases}$
Median	0
Mode	0
Variance	$\begin{cases} \frac{\nu}{\nu-2}, & \nu > 2 \\ \infty, & 1 < \nu \leq 2 \\ \text{undefined}, & \text{else} \end{cases}$
Skewness	$\begin{cases} 0, & \nu > 3 \\ \text{undefined}, & \text{else} \end{cases}$
Excess Kurtosis	$\begin{cases} \frac{6}{\nu-4}, & \nu > 4 \\ \infty, & 2 < \nu \leq 4 \\ \text{undefined}, & \text{else} \end{cases}$

5.7.3 Relation to Sample Mean

Let $X_1, \ldots, X_n$ be independent and identically distributed according to the distribution $\mathcal{N}(\mu, \sigma^2)$. Denote the sample mean

$$\overline{X} = \frac{1}{n}\sum_i X_i$$

and the sample variance

$$S^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2.$$

Then the random variable

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

follows the standard normal distribution $\mathcal{N}(0, 1)$. However the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

follows a Student's $t-$distribution with $\nu = n - 1$ degrees of freedom.

Notably, despite both being derived from the sample $X_1, \ldots, X_n$, the random variables respectively defined by the numerator and denominator of T are independent of one another.

5.8 *F*-Distribution

5.8.1 Interpretation

In statistics, the $F-$ distribution often arises as the null distribution of a test statistic, such as in an $F-$test.

5.8.2 Properties

Parameters	$d_1, d_2 \in \mathbb{N}$ degrees of freedom
Support	$x \in \begin{cases} (0,\infty), & d_1 = 1 \\ [0,\infty), & \text{else} \end{cases}$
Probability Density Function[6]	$\dfrac{\sqrt{\frac{(d_1x)^{d_1} d_2^{d_2}}{(d_1x+d_2)^{d_1+d_2}}}}{xB\left(\frac{d_1}{2},\frac{d_2}{2}\right)}$
Mean	$\dfrac{d_2}{d_2-2}$ for $d_2 > 2$
Mode	$\dfrac{d_1-2}{d_1} \cdot \dfrac{d_2}{d_2+2}$ for $d_1 > 2$
Variance	$\dfrac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$ for $d_2 > 4$
Skewness	$\dfrac{(2d_1+d_2-2)\sqrt{8(d_2-4)}}{(d_2-6)\sqrt{d_1(d_1+d_2-2)}}$ for $d_2 > 6$

5.8.3 Characterization

Let X_1 and X_2 be independent chi-square random variables with d_1 and d_2 degrees of freedom, respectively. Then the random variable defined by

$$F = \frac{U_1/d_1}{U_2/d_2}$$

follows an $F-$distribution with parameters d_1 and d_2. Equivalently, by the definition of the chi-square distribution, define F by

$$F = \frac{S_1^2/d_1\sigma_1^2}{S_2^2/d_2\sigma_2^2}$$

[6]Here, $B(x,y)$ denotes the beta function, $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$.

where S_1^2 and S_2^2 are the sum of d_1 and d_2 random variables following the normal distributions $\mathcal{N}(0, \sigma_1^2)$ and $\mathcal{N}(0, \sigma_2^2)$, respectively. In this case, F is also follow an $F-$distribution with parameters d_1 and d_2.

5.8.4 Related Distributions

- As above, if $X \sim \chi^2(d_1)$ and $Y \sim \chi^2(d_2)$, then $\dfrac{X/d_1}{Y/d_2} \sim F(d_1, d_2)$.
- If $X_k \sim \Gamma(\alpha_k, \beta_k)$, $i = 1, 2$ are independent, then $\dfrac{\alpha_2\beta_1X_1}{\alpha_1\beta_2X_2} \sim F(2\alpha_1, 2\alpha_2)$.
- If $X \sim \mathrm{Beta}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$ then $\dfrac{d_2X}{d_1(1-X)} \sim F(d_1, d_2)$. Equivalently, if $Y \sim F(d_1, d_2)$ then $\dfrac{d_1X/d_2}{1 + d_1X/d_2} \sim \mathrm{Beta}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$.
- If $X \sim F(d_1, d_2)$ then $X^{-1} \sim F(d_2, d_2)$.
- If $T \sim t(n)$ then $X^2 \sim F(1, n)$.

5.9 Maxwell-Boltzmann Distribution

5.9.1 Interpretation

The Maxwell-Boltzmann distribution models the speed of particles in an ideal gas. The speed is the norm of the velocity vector

$$V = ||\vec{V}|| = ||(V_x, V_y, V_z)|| = \sqrt{V_x^2 + V_y^2 + V_z^2}$$

and the velocity components are independent and identically distributed according to $\mathcal{N}(\mu, \sigma^2)$. Thus, we can apply an appropriate transformation so that mathematically, $V \sim \chi(3)$ with a scaling factor $a = \sqrt{kT/m}$ where T is the temperature of the gas mixture, m is the mass of the gas particles, and k is the Boltzmann constant.

5.9.2 Properties

Parameters	$a = \sqrt{kT/m} > 0$
Support	$x \in (0, \infty)$
Probability Density Function	$f(x) = \sqrt{\frac{2}{\pi}} \frac{x^2 e^{-\frac{-x^2}{2a^2}}}{a^3}$
Cumulative Distribution Function	$F(x) = \operatorname{erf}\left(\frac{x}{\sqrt{2}a}\right) - \sqrt{\frac{2}{\pi}} \frac{x e^{\frac{-x^2}{2a^2}}}{a}$
Mean	$2a\sqrt{\frac{2}{\pi}}$
Mode	$\sqrt{2}a$
Variance	$\frac{a^2(3\pi - 8)}{\pi}$
Skewness	$\frac{2\sqrt{2}(16 - 5\pi)}{(3\pi - 8)^{3/2}}$
Excess Kurtosis	$\frac{160\pi - 12\pi^2 - 384}{(3\pi - 8)^2}$

5.10 Beta Distribution

5.10.1 Interpretation

The beta distribution is used to model the behavior of random variables restricted to finite intervals, such as percentages and proportions. It is also the conjugate prior of the Bernoulli, binomial, negative binomial, and geometric distributions.

5.10.2 Properties

Notation	$\mathrm{Beta}(\alpha, \beta)$
Parameters	$\alpha, \beta > 0$
Support	$x \in [0,1]$ or $x \in (0,1)$
Probability Density Function[7]	$f(x) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$
Cumulative Distribution Function[8]	$F(x) = I_x(\alpha, \beta)$
Mean	$\dfrac{\alpha}{\alpha+\beta}$
Median	$\approx \dfrac{\alpha - \frac{1}{3}}{\alpha+\beta-\frac{2}{3}}$ for $\alpha, \beta > 1$
Mode	$\dfrac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$
Variance	$\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Skewness	$\dfrac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$
Excess Kurtosis	$\dfrac{6\left((\alpha-\beta)^2(\alpha+\beta+1)-\alpha\beta(\alpha+\beta+2)\right)}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}$

[8] Here, $I_x(\alpha, \beta)$ denotes the regularized incomplete beta function.

5.11 Logistic Distribution

5.11.1 Interpretation

The logistic distribution is the distribution defined by having a cumulative distribution function equal to the logistic function, which is commonly seen in logistic regression problems and feedforward neural networks.

5.11.2 Properties

Notation	Logistic(μ, s)
Parameters	$\mu \in \mathbb{R}$ $s > 0$
Support	$x \in \mathbb{R}$
Probability Density Function	$f(x) = \dfrac{e^{-\frac{x-\mu}{s}}}{s\left(1+e^{-\frac{x-\mu}{s}}\right)^2}$
Cumulative Distribution Function	$F(x) = \dfrac{1}{1+e^{-\frac{x-\mu}{s}}})$
$100Q^{th}$ Quantile	$\mu + s\ln\left(\dfrac{Q}{1-Q}\right)$
Mean	μ
Median	μ
Mode	μ
Variance	$\dfrac{s^2\pi^2}{3}$
Skewness	0
Excess Kurtosis	$\dfrac{6}{5}$
Entropy	$\ln s + 2$
Characteristic Function	$\varphi(t) = e^{i\mu t}\dfrac{st}{\sinh(\pi st)}$

5.12 Dirac Delta

Definition 5.9 The *Dirac delta function* is a functioned defined such that $\delta(x) = 0$ everywhere except $x = 0$ but with unit integral over the real line. That is

- $\delta(x) = \begin{cases} \infty, & x = 0 \\ 0, & x \neq 0 \end{cases}$
- $\int\limits_{-\infty}^{\infty} \delta(x)dx = 1.$

The Dirac delta is the continuous analog of the Kronecker delta.

5.12.1 Properties

- For any function $f(x)$,

$$\int\limits_{-\infty}^{\infty} f(x)\delta(x-a)dx = f(a).$$

- The Dirac delta function is the derivative of the Heaviside step function.

$$H(x) = \int\limits_{-\infty}^{x} \delta(t)dt = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

$$\delta(x) = \frac{d}{dx}H(x)$$

- For any $\alpha \in \mathbb{R}$,

$$\delta(\alpha x) = \frac{\delta(x)}{|\alpha|}.$$

- Let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable such that g' is nowhere zero, and suppose g has roots at $x_1, \ldots, x_n$. Then,

$$\delta \circ g(x) = \delta\left(g\left(x\right)\right) = \sum_i \frac{\delta(x - x_i)}{|g'(x_i)|}.$$

SECTION 6

MULTIVARIATE DISTRIBUTIONS

6.1 Joint Probability

Let $X_1, X_2, X_3, \ldots$ be random variables defined on a probability space $(\Omega, \mathcal{F}, P)$. The *joint probability distribution* is a probability distribution which describes the probability that each of the X_i takes on a given value or range of values. If there are only two X_i, this is called a *bivariate distrubtion*. In general this is called a *multivariate distribution*.

Definition 6.1 Let $X_1, \ldots, X_n$ be random variables. We can define a vector $\vec{X}$ using these by

$$\vec{X} = (X_1, \ldots, X_n)^T = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

In this case $\vec{X}$ is called a *random vector*.

Definition 6.2 Let $\vec{X} = (X_1, \ldots, X_n)^T$ be a random vector. The joint cumulative distribution function is the function

$$F_{\vec{X}}(\vec{x}) = P(X_1 \leq x_1, \ldots, X_n \leq x_n).$$

In the case when the X_i are independent, this reduces to

$$F_{\vec{X}}(\vec{x}) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n) = \prod_i P(X_i \leq x_i) = \prod_i F_{X_i}(x_i).$$

Definition 6.3 Let X and Y be independent discrete random variables. Their *joint probability mass function* is the function defined by

$$\begin{aligned} p_{X,Y}(x,y) &= P(X = x \text{ and } Y = y) \\ &= P(Y = y \mid X = x) \cdot P(X = x) \\ &= P(X = x \mid Y = y) \cdot P(Y = y) \end{aligned}$$

In general for a random vector $\vec{X} = (X_1, \ldots, X_n)^T$ the joint probability mass function is

$$\begin{aligned} p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 = x_1) \times P(X_2 = x_2 \mid X_1 = x_1) \times \cdots \\ \cdots \times P(X_n = x_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \end{aligned}$$

Definition 6.4 Let X and Y be independent continuous random variables. Their *joint probability density function* is the function defined by

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$$

From this, analogous to the discrete case,

$$f_{X,y}(x,y) = f_{Y|X}(y \mid x) f_X(x) = f_{X|Y}(x \mid y) f_Y(y).$$

In general, for a random vector $\vec{X} = (X_1, \ldots, X_n)^T$ the joint probability density function is

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{\partial^n F_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}$$

Again this gives

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) f_{X_2|X_1}(x_2 \mid x_1) \cdots \\ \cdots f_{X_n|X_1,\ldots,X_{n-1}}(x_n \mid x_1, \ldots, x_{n-1}) \end{aligned}$$

Definition 6.5 Let X and Y be a independent continuous and discrete random variables, respectively. Their *mixed joint density* is defined as

$$f_{X,Y}(x,y) = f_{X|Y}(x \mid y) P(Y = y) = P(Y = y \mid X = x) f_X(x).$$

This can be used to recover the joint cumulative distribution function

$$F_{X,Y}(x,y) = \sum_{t \leq y} \int_{s \leq x} f_{X,Y}(s,t) ds.$$

Definition 6.6 Let X and Y be two independent random variables with joint probability density $f_{X,Y}(x,y)$. The individual probability distributions of each random variable is referred two as its *marginal probability distribution*, and is given by

$$f_X(x) = \int f_{X,Y}(x,y)dy$$

and

$$f_Y(y) = \int f_{X,Y}(x,y)dx.$$

6.2 Dependence

Proposition 6.7 Let X and Y be random variables with probability density f_X and f_Y respectively and cumulative distribution F_X and F_Y respectively. Then

- X and Y are independent if and only if $F_{X,Y}(x,y) = F_X(x)F_Y(y)$.
- X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

Proposition 6.8 If a subset A of random variables $X_1, \ldots, X_n$ is conditionally dependent on another subset B of these variables, then

$$P(X_1, \ldots, X_n) = P(B) \cdot P(A \mid B)$$

and therefore can be represented by lower-dimensional probability distributions.

Definition 6.9 Let X and Y be random variables. The *covariance* of X and Y describes how these variables move together, and is defined as

$$\sigma_{XY} = \text{cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[XY] - \mu_X\mu_Y.$$

Definition 6.10 Let X and Y be random variables. The *correlation* of X is Y is the covariance of X and Y scaled by their respective standard deviations. The result is a unitless measure of how X and Y vary together, and is often easier to interpret than the covariance. It is denoted ρ_{XY} and is given by

$$\rho_{XY} = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

Two random variables whose correlation is nonzero, they are said to be *correlated*. Conversely if $\rho_{XY} = 0$ then X and Y are said to be *uncorrelated*.

Proposition 6.11 Let X, Y, W, and V be random variables and let $a, b \in \mathbb{R}$. Then

- If X and Y are independent then $\text{cov}(X, Y) = 0$. The converse is not true in general. For example, supposing X is uniformly distributed on $[-1, 1]$ and suppose further that $Y = X^2$. X and Y are clearly not independent but

  $$\text{cov}(X, Y) = \mathbb{E}[X \cdot X^2] - \mathbb{E}[X]\mathbb{E}[X^2] = \mathbb{E}[X^3] - 0 = \mathbb{E}]X^3] = 0.$$

- $\text{cov}(X, a) = 0$
- $\text{cov}(X, X) = \text{var}(X)$
- $\text{cov}(X, Y) = \text{cov}(Y, X)$
- $\text{cov}(aX, bY) = ab\text{cov}(X, Y)$
- $\text{cov}(X + a, Y + b) = \text{cov}(X, Y)$
- $\text{cov}(aX + bY, cW + dV) = ac\text{cov}(X, W) + ad\text{cov}(X, V) + bc\text{cov}(Y, W) + bd\text{cov}(Y, V)$
- For a random vector $\vec{X} = (X_1, \ldots, X_n)^T$ and $a_1, \ldots, a_n \in \mathbb{R}$,

  $$\begin{aligned} \text{var}\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n} a_i^2 \text{var}(X_i) + 2 \sum_{i<j} a_i a_j \text{cov}(X_i, X_j) \\ &= \sum_{i,j} a_i a_j \text{cov}(X_i, X_j) \end{aligned}$$

Theorem 6.12 – Hoeffding's Covariance Identity A useful way to compute the covariance between two random variables X and Y is

$$\text{cov}(X, Y) = \iint \left(F_{X,Y}(x, y) - F_X(x) F_Y(y)\right) dx dy.$$

6.3 Categorical Distribution

6.3.1 Interpretation

The categorical distribution is the generalization of the Bernoulli distribution to a random process where the outcome can be one of k categories each with an associated probability. For example, rolling a fair six-sided die follows the categorical distribution with six categories, each with probability $\frac{1}{6}$.

6.3.2 Properties

Parameters	$k \in \mathbb{N}$ categories $p_1, \dots, p_k > 0$, with $\sum p_i = 1$
Support	$x \in \{1, \dots, k\}$
Probability Mass Function	$f(x) = \mathbb{1}_{\{x=1\}} p_1 + \cdots + \mathbb{1}_{\{x=k\}} p_k$
Mode	i such that $p_i = \max\{p_1, \dots, p_k\}$

6.4 Multinomial Distribution

6.4.1 Interpretation

The multinomial distribution is the generalization of the binomial distribution to categorical processes over Bernoulli processes. For example, a multinomial could be used to model the number of times each number comes up in n rolls of a six-sided die.

6.4.2 Properties

Parameters	$n \in \mathbb{N}$ trials $p_1, \ldots, p_k > 0$ with $\sum p_i = 1$
Support	$x_i \in \{0, \ldots, n\}$ for $i \in \{1, \ldots, k\}$ such that $\sum x_i = n$
Probability Mass Function	$p(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$
Mean	$\mu_i = \mathbb{E}[X_i] = np_i$
Variance	$\mathrm{var}(X_i) = np_i(1 - p_i)$ $\mathrm{cov}(X_i, X_j) = -np_ip_j$ for $i \neq j$
Correlation	$\rho(X_i, X_j) = -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}$ for $i \neq j$

6.5 Multivariate Hypergeometric Distribution

6.5.1 Interpretation

The multivariate hypergeometric distribution extends the hypergeometric distribution to the case where there are n draws without replacement from a population of N objects which are subdivided into c categories of size K_i. On the other hand, the multinomial describes a similar situation but draws are taken with replacement.

6.5.2 Properties

Parameters	$c \in \mathbb{N}$ categories $(K_1, \ldots, K_c) \in \mathbb{N}^c$ with $\sum K_i = N$ $n = \{0, \ldots, N\}$ draws
Support	$(k_1, \ldots, k_c) \in \mathbb{N}^c$ with $\sum k_i = n$
Probability Mass Function	$p(\vec{k}) = \frac{\prod \binom{K_i}{k_i}}{\binom{N}{n}} = \frac{\binom{K_1}{k_1} \cdots \binom{K_c}{k_c}}{\binom{N}{n}}$
Mean	$\mu_i = \mathbb{E}[X_i] = n\frac{K_i}{N}$
Variance	$\mathrm{var}(X_i) = n\frac{N-n}{N-1}\frac{K_i}{N}\left(1 - \frac{K_i}{N}\right)$ $\mathrm{cov}(X_i, X_j) = -n\frac{N-n}{N-1}\frac{K_i}{N}\frac{K_j}{N}$
Moment Generating Function	$M(t) = e^{\vec{\mu}^T\vec{t} + \frac{1}{2}\vec{t}^T\Sigma\vec{t}}$
Characteristic Function	$\varphi(t) = e^{i\vec{\mu}^T\vec{t} - \frac{1}{2}\vec{t}^T\Sigma\vec{t}}$

6.6 Multivariate Normal Distribution

6.6.1 Interpretation

The multivariate normal distribution generalizes the normal distribution to higher dimensions. A random vector can be said to be k-variate normally distributed if each linear combination of its k component is normally distributed in the usual sense.

6.6.2 Properties

Notation	$\mathcal{N}(\vec{\mu}, \vec{\Sigma})$
Parameters	$\vec{\mu} = (\mu_1, \ldots, \mu_k) \in \mathbb{R}^k$ means $\Sigma \in \mathbb{R}^{k \times k}$, matrix of covariances
Support	$\vec{x} \in \mu + \operatorname{span}(\Sigma) \subseteq \mathbb{R}^k$
Probability Density Function	$f(\vec{x}) = -\frac{1}{\sqrt{2^k \pi^k \det \Sigma}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}$ exists only when Σ is positive definite
Mean	$\vec{\mu} = (\mu_1, \ldots, \mu_k)$
Mode	$\vec{\mu} = (\mu_1, \ldots, \mu_k)$
Variance	$\Sigma = [\sigma_{ij}] = [\operatorname{cov}(X_i, X_j)]$
Entropy	$\frac{1}{2} \ln \det(2\pi e \Sigma)$

Part II

Statistics

SECTION 7 OVERVIEW

Definition 7.1 In statistics, a collection of people, objects, processes, etc. under study is called the *population*.

Definition 7.2 A subset of the population which we study is called a *random sample*. These are often modeled as a set of random variables $X_1, \ldots, X_n$. The process of selecting a sample is called *sampling*.

Definition 7.3 The measured or observed results from a sample are called *data*. These are often modeled as values $x_1, \ldots, x_n$ of a random sample $X_1, \ldots, X_n$. For example one might consider $X_1, \ldots, X_n$ to be a Bernoulli trial modeling the results of a test for an infectious disease among the population. A singular observation is called a *datum*.

Definition 7.4 Data can either be *qualitative* or *quantitative*. Qualitative data, or *categorical data* is data which describes a sample by grouping each datum into different groups, whereas quantitative data describes a sample numerically. For example, hair color is a qualitative statistic whereas percentage of people with red hair is quantitative.

Definition 7.5 A *summary statistic* is a piece of information which summarizes a sample. For example one might consider the mean of the sample, $\overline{x} = \frac{1}{n}\sum x_i$. A *descriptive statistic* is a summary statistic which quantitatively describes a feature or features of a sample.

Definition 7.6 A *population parameter* is a numerical characteristic of the population at large which can be estimated via a descriptive statistic.

Definition 7.7 The number of times a specific value occurs in data is called its *frequency*. The *relative frequency* is the ratio of the number of times a value occurs to the total number of data. The *cumulative relative frequency* of a value is the sum of relative frequencies of all smaller (or equal) values.

Definition 7.8 An *experiment* is a process that is performed in order to investigate the relationship between two variables.

Definition 7.9 When one variable causes change in another, the first is called an *explanatory variable* while the second is called the *response variable*.

Definition 7.10 Any variable which is not an explanatory or response variable which can affect the outcome of a study is called a *lurking variable*

SECTION 8

ESTIMATING PARAMETERS

Definition 8.1 Given a random sample $X_1, \ldots, X_n$ of a population, the *sample mean* is denoted $\overline{X}$ and is defined as

$$\overline{X} = \frac{1}{n}\sum_{i=0}^{n} X_i$$

Definition 8.2 Given a random sample $X_1, \ldots, X_n$, the *Bessel corrected sample variance* is denoted S^2 and is given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Similarly the *sample standard deviation* is denoted S and is given by

$$S = \sqrt{S^2} = \sqrt{\frac{\sum(X_i - \overline{X})^2}{n-1}}.$$

Definition 8.3 Let $X_1, \ldots, X_n$ be a random sample. An *estimator* of a parameter θ is a transformation $\hat{\Theta}$ of the random sample. For example, the sample mean

$$\hat{\Theta}(X_1, \ldots, X_n) = \frac{X_1 + \cdots + X_n}{n}$$

is an estimator of the population mean μ.

An estimator is said to be *unbiased* if

$$\mathbb{E}(\hat{\Theta}) = \theta$$

and is *asymptotically unbiased* if

$$\mathbb{E}(\hat{\Theta}) \xrightarrow{n\to\infty} \theta.$$

Definition 8.4 Let $\hat{\Theta}$ be an estimator of the parameter θ. The difference $\mathbb{E}(\hat{\Theta}) - \theta$ is called the *bias* of the estimator. Evidently, the bias of an estimator is 0 if and only if the estimator is unbiased.

Definition 8.5 For a random sample $X_1, \ldots, X_n$, a *minimum variance unbiased estimator* is an unbiased estimator whose variance is less than or equal to the variance of any other possible unbiased estimator for the same parameter.

Theorem 8.6 – Cramér-Rao Lower Bound Let $\hat{\Theta}$ be an unbiased estimator of a deterministic (fixed but unknown) parameter θ. Then

$$\operatorname{Var}(\hat{\Theta}) \geq \frac{1}{n\mathbb{E}\left[\left(\dfrac{\partial \ln f(x \mid \theta)}{\partial\theta}\right)^2\right]} \equiv \frac{1}{-n\mathbb{E}\left[\dfrac{\partial^2 \ln f(x \mid \theta)}{\partial\theta^2}\right]}$$

Definition 8.7 Let $X_1, \ldots, X_n$ be a random sample. $\hat{\Phi}(X_1, \ldots, X_n)$ is called a *sufficient statistic* for θ if the joint density function of the sample $\prod_{i=1}^{n} f(x_i \mid \theta)$ can be factored into a product of a function of θ and $\hat{\Phi}$ only, and a function of $x_1, \ldots, x_n$ only:

$$\prod_{i=1}^{n} f(x_i \mid \theta) = g(\theta, \hat{\Phi}) \cdot h(x_1, \ldots, x_n)$$

For example, if the X_i are Bernoulli distributed with parameter p, then

$$\prod_{i=1}^{n} f(x_i \mid p) = p^{x_1+\cdots+x_n}(1-p)^{n-x_1-\cdots-x_n}$$

so defining $\hat{\Phi}(X_1, \ldots, X_n) = \sum X_i$,

$$\begin{aligned}\prod_{i=1}^{n} f(x_i \mid p) &= p^{\hat{\Phi}}(1-p)^{n-\hat{\Phi}} \\ &= g(p, \hat{\Phi})\end{aligned}$$

so $\hat{\Phi}$ is a sufficient statistic for p. To make $\hat{\Phi}$ an unbiased estimator for p, we define

$$\hat{\Theta}(X_1, \ldots, X_n) = \frac{\hat{\Phi}(X_1, \ldots, X_n)}{n} = \frac{X_1 + \cdots + X_n}{n}$$

as would be expected.

8.1 Method of Moments

The method of moments is the easier of the two commons ways to determine parameter estimators. It provides good estimators in many cases but can sometimes result in inefficient estimators.

Suppose we need to estimate parameters $\theta_1, \ldots, \theta_k$ of a random sample $X_1, \ldots, X_n$. Then we express the first k moments of X as functions of the θ_i:

$$\begin{aligned}
\mu_1 &= \mathbb{E}[X] = g_1(\theta_1, \ldots, \theta_k) \\
\mu_2 &= \mathbb{E}[X^2] = g_2(\theta_1, \ldots, \theta_k) \\
&\vdots \\
\mu_k &= \mathbb{E}[X^k] = g_k(\theta_1, \ldots, \theta_k)
\end{aligned}$$

and where possible, invert this system of equations to express

$$\begin{aligned}
\hat{\Theta}_1 &= h_1(\hat{\mu}_1, \ldots, \hat{\mu}_k) \\
\hat{\Theta}_2 &= h_2(\hat{\mu}_1, \ldots, \hat{\mu}_k) \\
&\vdots \\
\hat{\Theta}_k &= h_k(\hat{\mu}_1, \ldots, \hat{\mu}_k)
\end{aligned}$$

Example 8.8 Suppose $X_1, \ldots, X_n$ is a normally distributed random sample. We need estimators $\hat{\mu}$ and $\hat{\sigma}^2$ for the mean and variance respectively. We have

$$\mathbb{E}[X] = \hat{\mu}$$

and

$$\mathbb{E}[X^2] = \hat{\sigma}^2 + \hat{\mu}^2$$

which together suggest that

$$\hat{\mu} = \mathbb{E}[X] = \frac{\sum X_i}{n} = \overline{X}$$

and

$$\hat{\sigma}^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{n}$$

This $\hat{\sigma}^2$ is not unbiased, although it is asymptotically unbiased and it can be made into an unbiased estimator easily.

8.2 Maximum Likelihood Estimation

This method will typically find a minimum variance unbiased estimator, though it may sometimes produce only an asymptotically unbiased estimator. This has the drawback that the estimators will sometimes turn out to be very complicated functions of the X_i.

Definition 8.9 Suppose we have a random sample $X_1, \ldots, X_n$ and want to estimate parameters $\theta_1, \ldots, \theta_k$. In the joint density function $\prod f(x_i \mid \theta_1, \ldots, \theta_n)$, replace the x_i with there observed values and turn it into a function $f_{x_i}(\theta_1, \ldots, \theta_n)$ of the parameters. This is called the *likelihood function*. To obtain the estimators of the θ_j, maximize the likelihood function f_{x_i}. The corresponding optimal values of θ_j are the parameter estimates.

As $\ln x$ is a continuously differentiable monotonic increasing function, maximizing

$$\prod_i f_{x_i}(\theta_1, \ldots, \theta_n)$$

is equivalent to maximizing

$$\ln\left(\prod_i f_{x_i}(\theta_1, \ldots, \theta_n)\right) = \sum_i \ln f_{x_i}(\theta_1, \ldots, \theta_n)$$

Example 8.10 Suppose $X_1, \ldots, X_n$ is a geometrically distributed random sample. Then

$$\begin{aligned}
\prod_i f_{x_i}(p) &= \prod_i (1-p)^{x_i} p \\
&= p^n (1-p)^{\sum_i x_i} \\
\ln\left(\prod_i f_{x_i}(p)\right) &= n \ln p + \ln(1-p) \sum_i x_i
\end{aligned}$$

Thus to maximize the likelihood function,

$$\begin{aligned}\frac{d \ln \prod_i f_{x_i}(p)}{dp} &= \frac{n}{p} - \frac{\sum x_i}{1-p} \\ 0 &= \frac{n}{p} - \frac{\sum x_i}{1-p} \\ \Rightarrow \hat{p} &= \frac{n}{n + \sum x_i}\end{aligned}$$

8.3 Confidence Intervals

Definition 8.11 Rather than give a point estimate for a parameter, which can in essence never be correct, it is often helpful to talk about an interval which has some nonzero chance of containing the correct value. Thus for a parameter θ and an estimate of that parameter $\hat{\theta}$, we might say that θ falls in the interval $\hat{\theta} \pm w$ or $[\hat{\theta} - w, \hat{\theta} + w]$. This interval is called a *confidence interval* for the parameter θ.

Definition 8.12 We will typically associate to a confidence interval a *level of confidence* for that interval. This is usually expressed as a value $1 - \alpha$.

Remark

In this context, $1 - \alpha$ is **not** the probability that the true value of θ is contained in the obtained interval. This value is deterministic and already exists external to the test, and thus it does not make sense to talk about it in a probabilistic sense. Rather, it is the a priori probability that an interval obtained by the test will contain the true value of the parameter. That is, before any measurements are taken, it is the probability that over all possible data sets, a data set is measured which produces a confidence interval which is representative of the true value.

8.3.1 For the mean μ with known σ

Let $X_1, \ldots, X_n$ be a random sample and suppose that the standard deviation σ is known but the mean μ is not. As n becomes large,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

resembles a standard normal distribution, and thus

$$\begin{aligned} P(|Z| < z_{\alpha/2}) &= P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ &= 2\Phi(z_{\alpha/2}) - 1 \\ &= 1 - \alpha \end{aligned}$$

Consequently, this identifies the interval

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

as a $100(1-\alpha)\%$ confidence interval for the parameter μ.

Remark

A common choice for $\alpha = 0.05$. Note that in general $P(Z < z_\alpha) = 1 - \alpha$.

8.3.2 For the mean μ with unknown σ

Let $X_1, \ldots, X_n$ be a random sample and suppose that the standard deviation σ and the meanμ are both unknown. In this case we replace the σ above with the sample standard deviation s. Then the random variable T defined by

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

follows a Student's $t-$distribution with $n-1$ degrees of freedom rather than the standard normal distribution. Thus by a similar argument

$$\begin{aligned} P(|T| < t_{\alpha/2}(n-1)) &= P\left(\overline{X} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right) \\ &= 1 - \alpha \end{aligned}$$

Consequently, this identifies the interval

$$\overline{X} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} = \left[\overline{X} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right]$$

as a $100(1-\alpha)\%$ confidence interval for the parameter μ.

Remark

When n is very large, say ≥ 30, there is little difference between $z_{\alpha/2}$ and $t_{\alpha/2}(n-1)$. We could use $z_{\alpha/2}$ rather than $t_{\alpha/2}(n-1)$ in this case.

8.3.3 Difference of two means with known σ

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent random samples with unknown means μ_X and μ_Y and with the same known standard deviation σ. Then as n and m become large the random variable

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$$

follows a standard normal distribution. This identifies

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

or equivalently

$$\left[\overline{X} - \overline{Y} - z_{\alpha/2}\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}, \overline{X} - \overline{Y} + z_{\alpha/2}\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}\right]$$

as a $100(1-\alpha)\%$ confidence interval for the difference between μ_X and μ_Y.

8.3.4 Difference of two means with unknown σ

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent random samples with unknown means μ_X and μ_Y and same but unknown standard deviation.

Definition 8.13 The *pooled sample standard deviation* is defined to be

$$s_p = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}}$$

where s_X and s_Y are the sample standard deviations of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ respectively.

Then the variable T defined by

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

follows a Student's $t-$distribution with $n+m-2$ degrees of freedom. This identifies

$$(\overline{X} - \overline{Y}) \pm t_{\alpha/2}(n+m-2)s_p\sqrt{\frac{1}{n} + \frac{1}{m}}$$

or equivalently

$$\left[\overline{X} - \overline{Y} - t_{\alpha/2}(n+m-2)s_p\sqrt{\frac{1}{n} + \frac{1}{m}}, \overline{X} - \overline{Y} + t_{\alpha/2}(n+m-2)s_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right]$$

as a $100(1-\alpha)\%$ confidence interval for the difference between μ_X and μ_Y.

8.3.5 Difference of two means with known σ_X and σ_Y

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent random samples with unknown means μ_X and μ_Y and known but different standard deviations σ_X and σ_Y. Then as n and m become large, the random variable

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

follows a standard normal distribution $\mathcal{N}(0,1)$. This identifies

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

or equivalently

$$\left[\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right]$$

as a $100(1-\alpha)\%$ confidence interval for the difference between μ_X and μ_Y.

8.3.6 Difference of two means with unknown $\boldsymbol{\sigma_X}$ and $\boldsymbol{\sigma_Y}$

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent random variables with unknown means μ_X and μ_Y as well as unknown standard deviations σ_X and σ_Y. In this case we are not able to construct a random variable from these random samples with a simple distribution. However if n and m are both large, say $n, m \geq 30$ or so, then we can replace σ_X and σ_Y with s_X and s_Y respectively, where s_X is the sample standard deviation of X and similar for σ_Y. In the case of large n and m, the random variable

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

is approximately standard normally distributed. Thus we can identify an approximate $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ as

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

or equivalently

$$\left[\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}, \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}\right]$$

8.3.7 For proportions

Let $X_1, \ldots, X_n$ be a random sample of Bernoulli trials with unknown parameter p. This can be thought of as a test of the population for some special or distinguishing feature. For example this could be the results of a test of citizens for an infectious disease.

Definition 8.14 As the X_i can take on a value of 1 (when the test subject has the specific feature being tested for) or 0 (when they don't), the *sample proportion.* of cases is equal to the sample mean

$$\hat{p} = \overline{X} = \frac{X_1 + \cdots + X_n}{n}.$$

Moreover, in the case when n is large, it follows that $\hat{p}$ is approximately normally distributed with mean p and standard deviation $\frac{p(1-p)}{n}$. It follows that

$$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}}$$

is approximately standard normally distributed. It follows that

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

is an approximate $100(1-\alpha)\%$ confidence interval for p.

8.3.8 Difference of proportions

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent random samples of Bernoulli trials with unknown parameters p_X and p_Y. When n is large, the random variable

$$Z = \frac{(\hat{p}_X - \hat{p}_Y) - (p_X - p_Y)}{\sqrt{\dfrac{\hat{p}_X(1-\hat{p}_X)}{n} + \dfrac{\hat{p}_Y(1-\hat{p}_Y)}{m}}}$$

is approximately standard normally distributed. It follows that

$$(\hat{p}_X - \hat{p}_Y) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{m}}$$

or equivalently

$$\left[\hat{p}_X - \hat{p}_Y - z_{\alpha/2}\sqrt{\tfrac{\hat{p}_X(1-\hat{p}_X)}{n} + \tfrac{\hat{p}_Y(1-\hat{p}_Y)}{m}}, \hat{p}_X - \hat{p}_Y + z_{\alpha/2}\sqrt{\tfrac{\hat{p}_X(1-\hat{p}_X)}{n} + \tfrac{\hat{p}_Y(1-\hat{p}_Y)}{m}}\right]$$

is an approximate $100(1-\alpha)\%$ confidence interval for the difference between p_X and p_Y.

8.3.9 For variances

Let $X_1, \dots, X_n$ be a normally distributed random sample with unknown variance σ^2. Then the random variable

$$X = \frac{(n-1)s^2}{\sigma^2}$$

where s^2 is the sample variance follows a chi-square distribution with $n-1$ degrees of freedom, $\chi^2(n-1)$. It follows that

$$P\left(\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1)\right) = 1-\alpha$$

and consequently the interval

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\right]$$

is a $100(1-\alpha)\%$ confidence interval for σ^2.

8.3.10 For ratios of variances

Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ be independent normally distributed random variables with unknown variances σ_X^2 and σ_Y^2 respectively. Then the random variable F defined by

$$F = \frac{\frac{s_X^2}{\sigma_X^2}}{\frac{s_Y^2}{\sigma_Y^2}}$$

follows an F distribution with parameters $n-1$ and $m-1$, $F_{n-1,m-1}$. Thus

$$P\left(F_{1-\alpha/2}(n-1,m-1) < \frac{s_X^2\sigma_Y^2}{s_Y^2\sigma_X^2} < F_{\alpha/2}(n-1,m-1)\right) = 1-\alpha$$

and consequently

$$\left[\frac{1}{F_{\alpha/2}(n-1,m-1)}\frac{s_X^2}{s_Y^2}, \frac{s_X^2}{s_Y^2}F_{\alpha/2}(n-1,m-1)\right]$$

is a $100(1-\alpha)\%$ confidence interval for the ratio $\frac{\sigma_X^2}{\sigma_Y^2}$.

SECTION 9 HYPOTHESIS TESTING

9.1 Overview

Definition 9.1 In order to perform a statistical test we first define a default hypothesis to test against. This is called the *null hypothesis* and is denoted H_0. The alternative to the null hypothesis is called the *alternate hypothesis* and is denoted H_A. If sufficient evidence is found to show that the null hypothesis does not hold, then we *reject* and *accept* the alternate hypothesis. If sufficient evidence is not found to show this, then we say we *fail to reject* the null hypothesis.

Definition 9.2 Part of defining a test is deciding what counts as the aforementioned "sufficient evidence" to reject the null hypothesis. To do this, we select a *significance level*, usually denoted α. The significance level allows us to determine how selective the test is. The higher the significance level, the easier it is to reject the null hypothesis, and vice versa. The significance level determines the acceptable probability of rejecting the null hypothesis when it is true. Common choices for α include 0.1, 0.05, 0.02, and 0.01. However, this is just convention.

Remark

Failing to reject the null hypothesis does not mean that the null hypothesis is accepted nor that it is true. It simply means that we have not found sufficient evidence to accept the alternate hypothesis.

Definition 9.3 There are two types of errors we can make when testing hypotheses. These are

- Rejecting H_0 when it is true – a *type I error*
- Accepting H_0 when it is false – a *type II error*

A type I error happens with probability α by definition. The probability of a type II error is typically denoted β and depends on the actual values of the parameters involved. The more common quantity to look at is $1 - \beta$, which is called the *power function* of the test .

Definition 9.4 Commonly we talk about a null hypothesis wherein a parameter is equal to a specific value $H_0 : \theta = \theta_0$, and an alternative hypothesis encompassing every other scenario, $H_A : \theta \neq \theta_0$. Such a test is said to be *two-sided*. This will often involve constructing a $100(1 - \alpha)\%$ interval $(q_{\alpha/2}, q_{1-\alpha/2}]$, where $q_{\alpha/2}$ is the $100(\alpha/2)^{th}$ quantile of the distribution.

We may instead consider the null hypothesis $H_0 : \theta \geq \theta_0$ and the alternate hypothesis $H_A : \theta < \theta_0$, or vice-versa. Such a test is said to be *one-sided*. This will instead involve constructing a $100(1 - \alpha)\%$ interval as (q_α, ∞) or $(-\infty, q_{1-\alpha})$ as appropriate.

Definition 9.5 Suppose we have selected a statistical test, and this produces a confidence interval I as shown above. The the region $\mathbb{R} \setminus I$ is called the *critical region* of the test. The critical region is exactly the values of the test statistic which, if observed, will result in the rejection of the null hypothesis.

Definition 9.6 The *p−value* of a statistical test represents the probability of the test statistic X achieving a value at least as extreme as the observed value simply due to random chance. That is,

- $p = P(X \leq x \mid H_0)$ for a one-sided right tail test
- $p = P(X \leq\mid H_0)$ for a one-sided left tail test
- $p = 2\min\{P(X \leq x \mid H_0), P(X \geq x \mid H_0)\}$ for a two-sided test

Remark

The p-value does not represent the probability that the null hypothesis is true or that the alternative hypothesis is false. It is not the probability that the observed effects were produced by random chance alone. It is not indicative of the overall size or importance of the observed effect.

The p-value is the calculated probability that the test statistic could be as extreme as it was observed to be under the assumption that the null hypothesis is true. As such, it is more of a statement about the relationship of the observed data to the hypothesis.

9.2 Statistical Testing Process

This is a typical timeline of events for performing a statistical test.

1. There is a research hypothesis. From this, formulate and state the null and alternative hypotheses.
2. Consider the statistical assumptions being made about the data in the test.
3. Decide on an appropriate test and state the relevant test statistic.
4. Derive the distribution of the test statistic under the null hypothesis. Typically this will be well known.
5. Select a significance level, α, a probability threshold below which the null hypothesis will be rejected. Common choices include 5% and 1%.
6. Using the distribution of the test statistic, determine the critical region – the range of values of the statistic for which the null hypothesis will be rejected.
7. Compute the observed value of the test statistic.
8. If the observed value is in the critical region, reject H_0. Otherwise, fail to reject H_0.

we could also replace steps **??** through **??** with the following.

6. Compute the observed value of the test statistic.
7. Determine the $p-$value.
8. Reject the null hypothesis in favour of the alternate hypothesis if and only if the $p-$value is less than or equal to the significance level α.

9.3 Common Parameter Tests

Tests for means			
Assumption	H_0	Test statistic T	Distribution of T
Normal population, σ known	$\mu = \mu_0$	$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\mathcal{N}(0,1)$
Normal population, σ unknown	$\mu = \mu_0$	$\frac{\overline{X} - \mu_0}{s/\sqrt{n}}$	$t(n-1)$
Any population, large n, σ unknown	$\mu = \mu_0$	$\frac{\overline{X} - \mu_0}{s/\sqrt{n}}$	$\approx \mathcal{N}(0,1)$
Two normal populations, same unknown σ	$\mu_X = \mu_Y$	$\frac{\overline{X} - \overline{Y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$	t_{n+m-2}

Tests for variance			
Assumption	H_0	Test statistic T	Distribution of T
Normal population	$\sigma = \sigma_0$	$\frac{(n-1)s^2}{\sigma^2}$	$\chi^2(n-1)$
Two normal populations	$\sigma_X = \sigma_Y$	$\frac{s_X^2\sigma_Y^2}{s_Y^2\sigma_X^2}$	$F(n-1, m-1)$

Concerning proportions			
Assumption	H_0	Test statistic T	Distribution of T
One population, large n	$p = p_0$	$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$	$\approx \mathcal{N}(0,1)$
k populations, large n_i	$p_1 = p_2 = \cdots = p_k$	$\frac{\sum n_i(\hat{p}_i - \hat{\hat{p}})^2}{\hat{\hat{p}}(1-\hat{\hat{p}})}$	$\approx \chi^2(k-1)$

9.4 Chi-Square Independence Test

Let $X_1, \ldots, X_n$ be a random sample with two nominal categorical variables denoted C_i and D_j with $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, \ell\}$. Let o_{ij} represent the number of observations in both category C_i and category D_j. It follows that $\sum\limits_j o_{ij}$ is the total number of observations in category C_i, $\sum\limits_i o_{ij}$ is the total number of observations in category D_j, and $\sum\limits_i \sum\limits_j o_{ij}$ is the total number of observations. This can be laid out in a table or matrix as follows.

	D_1	$\cdots$	D_j	$\cdots$	D_ℓ	Total
C_1	o_{11}	$\cdots$	o_{1j}	$\cdots$	$o_{1\ell}$	$\sum_j o_{1j}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\iddots$	$\vdots$	$\vdots$
C_i	o_{i1}	$\cdots$	o_{ij}	$\cdots$	$o_{i\ell}$	$\sum_j o_{ij}$
$\vdots$	$\vdots$	$\iddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_k	o_{k1}	$\cdots$	o_{kj}	$\cdots$	$o_{k\ell}$	$\sum_j o_{kj}$
Total	$\sum_i o_{i1}$	$\cdots$	$\sum_i o_{ij}$	$\cdots$	$\sum_i o_{i\ell}$	$\sum_i \sum_j o_{ij}$

Definition 9.7 The above table is called a *contingency table*. The individual o_{ij} are known as *observed frequencies*

Definition 9.8 We define a quantity known as the *expected frequency* for each pair i, j as follows.

$$e_{ij} = \frac{\left(\sum_j o_{ij}\right)\left(\sum_i o_{ij}\right)}{\sum_i \sum_j o_{ij}}$$

Remark

To perform this test, we will generally require that each e_{ij} be no less than 5.

We will define a test statistic

$$X = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

which, under the the null hypothesis H_0 that the categorical variables are independent, follows the distribution $\chi^2\left((k-1)(\ell-1)\right)$. If the observed value of X is too extreme as defined by the p value of the test, we reject the null hypothesis and declare the categories not to be independent.

9.5 Pearson's Chi-Square Goodness of Fit Test

Let $X_1, \ldots, X_n$ be a random sample of unknown distribution, and suppose we would like to test how well an arbitrary distribution with cumulative distribution function F fits the observed data. To do this, we first section the data into discrete groups $(x_i, x_{i+1}]$, for $i \in \{1, \ldots, m\}$, such that

$$-\infty = x_1 < x_2 < \cdots < x_i < x_{i+1} < \cdots < x_m < x_{m+1} = \infty$$

For each i, define the expected frequency for the i^{th} interval to be

$$E_i = \left(F\left(x_{i+1}\right) - F\left(x_i\right)\right) n$$

Define a test statistic

$$T = \sum_{i=1}^{n} \frac{\left(O_i - E_i\right)^2}{E_i}$$

where O_i is the number of observations falling in the i^{th} interval, $(x_i, x_{i+1}]$.

The variable T, under the null hypothesis H_0, follows a chi-square distribution with $(k - c)$ degrees of freedom, where k is the number of intervals which contain one or more observations and c is the number of parameters that were estimated from the observations (for example, the mean or the variance) plus one. For example, if we were testing the goodness of fit of a normal distribution with unknown mean and variance, $c = 3$.

If the observed value of T is too extreme as defined by the p value of the test, we reject the null hypothesis and declare that the random sample does not follow the proposed distribution.

SECTION 10
ORDER STATISTICS

Definition 10.1 Let $X_1, \ldots, X_n$ be a random sample. Define a new set of random variables $X_{(1)}, \ldots, X_{(n)}$ defined such that for each k, $X_{(k)}$ is the k^{th} smallest value among the X_i. $X_{(k)}$ is called the k^{th} *order statistic*. Note that when n is odd, $X_{\left(\frac{n+1}{2}\right)}$ is the sample median.

Remark 10.2 To find the marginal density function of $X_{(k)}$ for some k, we have

$$\begin{aligned} f_{(i)}(x) &= \lim_{\Delta \to 0} \frac{P(x \leq X_{(k)} < x + \Delta)}{\Delta} \\ &= \lim_{\Delta \to 0} \binom{n}{i-1, 1, n-i} F(x)^{i-1} \left(\frac{F(x+\Delta) - F(x)}{\Delta} (1 - F(x+\Delta)) \right)^{n-i} \\ &= \frac{n!}{(i-1)!(n-i)!} F(x)^{i-1} \left[1 - F(x)\right]^{n-i} f(x) \end{aligned}$$

Remark 10.3 The joint distribution of two order statistics $X_{(i)}$ and $X_{(j)}$ $(i < j)$ can be determined similarly as

$$f(x, y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} F(x)^{i-1} f(x) \left[F(y) - F(x)\right]^{j-i-1} f(y) \left[1 - F(y)\right]^{n-j}$$

for $x < y$ in the support of the distribution of the X_i.

SECTION 11 REGRESSION

Regression is a process by which the effects of one or more random variables on another random variable are described.

Definition 11.1 An *explanatory variable* is a variable whose impact on another variable is being studied. It is also called an *independent variable* or a *regressor variable.*

Definition 11.2 A *response variable* is a variable whose response to one or more explanatory variables is being studied. It is also called a *dependent variables.*

11.1 Simple Linear Regression

Definition 11.3 The *regression model* for simple linear regression is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$$

and assumes that

- There is no error on the x_i, and
- The error term ϵ is normally distributed according to $\mathcal{N}(0, \sigma^2)$.

Part III

Tables

Table ?? - Binomial Coefficients

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \binom{n}{n-r}$$

n	$\binom{n}{0}$	$\binom{n}{1}$	$\binom{n}{2}$	$\binom{n}{3}$	$\binom{n}{4}$	$\binom{n}{5}$	$\binom{n}{6}$	$\binom{n}{7}$	$\binom{n}{8}$	$\binom{n}{9}$	$\binom{n}{10}$	$\binom{n}{11}$	$\binom{n}{12}$
0	1												
1	1	1											
2	1	2	1										
3	1	3	3	1									
4	1	4	6	4	1								
5	1	5	10	10	5	1							
6	1	6	15	20	15	6	1						
7	1	7	21	35	35	21	7	1					
8	1	8	28	56	70	56	28	8	1				
9	1	9	36	84	126	126	84	36	9	1			
10	1	10	45	120	210	252	210	120	45	10	1		
11	1	11	55	165	330	462	462	330	165	55	11	1	
12	1	12	66	220	495	792	924	792	495	220	66	12	1
13	1	13	78	286	715	1287	1716	1716	1287	715	286	78	13
14	1	14	91	364	1001	2002	3003	6435	3432	2002	1001	364	91
15	1	15	105	455	1365	3003	5005	6435	6435	5005	3003	1365	455

Table ?? - Binomial Distribution

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$$

		p									
n	x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.4	0.45	0.5
2	0	0.902	0.810	0.722	0.640	0.563	0.490	0.423	0.360	0.302	0.250
	1	0.998	0.990	0.978	0.960	0.938	0.910	0.878	0.840	0.798	0.750
3	0	0.857	0.729	0.614	0.512	0.422	0.343	0.275	0.216	0.166	0.125
	1	0.993	0.972	0.939	0.896	0.844	0.784	0.718	0.648	0.575	0.500
	2	1.000	0.999	0.997	0.992	0.984	0.973	0.957	0.936	0.909	0.875
4	0	0.815	0.656	0.522	0.410	0.316	0.240	0.179	0.130	0.092	0.062
	1	0.986	0.948	0.890	0.819	0.738	0.652	0.563	0.475	0.391	0.313
	2	1.000	0.996	0.988	0.973	0.949	0.916	0.874	0.821	0.759	0.688
	3	1.000	1.000	0.999	0.998	0.996	0.992	0.985	0.974	0.959	0.938
5	0	0.774	0.590	0.444	0.328	0.237	0.168	0.116	0.078	0.050	0.031
	1	0.977	0.919	0.835	0.737	0.633	0.528	0.428	0.337	0.256	0.187
	2	0.999	0.991	0.973	0.942	0.896	0.837	0.765	0.683	0.593	0.500
	3	1.000	1.000	0.998	0.993	0.984	0.969	0.946	0.913	0.869	0.812
	4	1.000	1.000	1.000	1.000	0.999	0.998	0.995	0.990	0.982	0.969
6	0	0.735	0.531	0.377	0.262	0.178	0.118	0.075	0.047	0.028	0.016
	1	0.967	0.886	0.776	0.655	0.534	0.420	0.319	0.233	0.164	0.109
	2	0.998	0.984	0.953	0.901	0.831	0.744	0.647	0.544	0.442	0.344
	3	1.000	0.999	0.994	0.983	0.962	0.930	0.883	0.821	0.745	0.656
	4	1.000	1.000	1.000	0.998	0.995	0.989	0.978	0.959	0.931	0.891
	5	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.996	0.992	0.984
7	0	0.698	0.478	0.321	0.210	0.133	0.082	0.049	0.028	0.015	0.008
	1	0.956	0.850	0.717	0.577	0.445	0.329	0.234	0.159	0.102	0.063
	2	0.996	0.974	0.926	0.852	0.756	0.647	0.532	0.420	0.316	0.227
	3	1.000	0.997	0.988	0.967	0.929	0.874	0.800	0.710	0.608	0.500
	4	1.000	1.000	0.999	0.995	0.987	0.971	0.944	0.904	0.847	0.773
	5	1.000	1.000	1.000	1.000	0.999	0.996	0.991	0.981	0.964	0.938
	6	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.996	0.992
8	0	0.663	0.430	0.272	0.168	0.100	0.058	0.032	0.017	0.008	0.004
	1	0.943	0.813	0.657	0.503	0.367	0.255	0.169	0.106	0.063	0.035
	2	0.994	0.962	0.895	0.797	0.679	0.552	0.428	0.315	0.220	0.145
	3	1.000	0.995	0.979	0.944	0.886	0.806	0.706	0.594	0.477	0.363
	4	1.000	1.000	0.997	0.990	0.973	0.942	0.894	0.826	0.740	0.637
	5	1.000	1.000	1.000	0.999	0.996	0.989	0.975	0.950	0.912	0.855
	6	1.000	1.000	1.000	1.000	1.000	0.999	0.996	0.991	0.982	0.965
	7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.996

Table ?? - Poisson Distribution

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} \frac{\lambda^k e^{-\lambda}}{k!}$$

	$\lambda = \mathbb{E}[X]$									
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.905	0.819	0.741	0.670	0.607	0.549	0.497	0.449	0.407	0.638
1	0.995	0.982	0.963	0.938	0.910	0.878	0.844	0.809	0.772	0.736
2	1.000	0.999	0.996	0.992	0.986	0.977	0.966	0.953	0.937	0.920
3	1.000	1.000	1.000	0.999	0.998	0.997	0.994	0.991	0.987	0.981
4	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.996
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999

x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0.333	0.301	0.273	0.247	0.223	0.202	0.183	0.165	0.150	0.135
1	0.699	0.663	0.627	0.592	0.558	0.525	0.493	0.463	0.434	0.406
2	0.900	0.879	0.857	0.833	0.809	0.783	0.757	0.731	0.704	0.677
3	0.974	0.966	0.957	0.946	0.934	0.921	0.907	0.891	0.875	0.857
4	0.995	0.992	0.989	0.986	0.981	0.976	0.970	0.964	0.956	0.947
5	0.999	0.998	0.998	0.997	0.996	0.994	0.992	0.990	0.987	0.983
6	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.997	0.995
7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999

x	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0
0	0.111	0.091	0.074	0.061	0.050	0.041	0.033	0.027	0.022	0.018
1	0.355	0.308	0.267	0.231	0.199	0.171	0.147	0.126	0.107	0.092
2	0.623	0.570	0.518	0.469	0.423	0.380	0.340	0.303	0.269	0.238
3	0.819	0.779	0.736	0.692	0.647	0.603	0.558	0.515	0.473	0.433
4	0.928	0.904	0.877	0.848	0.815	0.781	0.744	0.706	0.668	0.629
5	0.975	0.964	0.951	0.935	0.916	0.895	0.871	0.844	0.816	0.785
6	0.993	0.988	0.983	0.976	0.966	0.955	0.942	0.927	0.909	0.889
7	0.998	0.997	0.995	0.992	0.988	0.983	0.977	0.969	0.960	0.949
8	1.000	0.999	0.999	0.998	0.996	0.994	0.992	0.988	0.984	0.979
9	1.000	1.000	1.000	0.999	0.999	0.998	0.997	0.996	0.994	0.992
10	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.997
11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999

Table ?? - Chi-Square Distribution

$$F(x) = P(X \le x) = \int_0^x \frac{1}{2^{k/2}\Gamma(k/2)} t^{\frac{k}{2}-1} e^{-t/2} dt$$

	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
k	$\chi^2_{0.99}(k)$	$\chi^2_{0.975}(k)$	$\chi^2_{0.95}(k)$	$\chi^2_{0.90}(k)$	$\chi^2_{0.10}(k)$	$\chi^2_{0.05}(k)$	$\chi^2_{0.025}(k)$	$\chi^2_{0.01}(k)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21
11	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72
12	3.571	4.404	5.226	6.304	18.55	21.03	23.34	26.22
13	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69
14	4.660	5.629	6.571	7.790	21.06	23.68	26.12	29.14
15	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58
16	5.812	6.908	7.962	9.312	23.54	26.30	28.84	32.00
17	6.408	7.564	8.672	10.08	24.77	27.59	30.19	33.41
18	7.015	8.231	9.390	10.86	25.99	28.87	31.53	34.80
19	7.633	8.907	10.12	11.65	27.20	30.14	32.85	36.19
20	8.260	9.591	10.85	12.44	28.41	31.41	34.17	37.57
21	8.897	10.28	11.59	13.24	29.62	32.67	35.48	38.93
22	9.542	10.98	12.34	14.04	30.81	33.92	36.78	40.29
23	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64
24	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
40	22.16	24.43	26.51	29.05	51.80	55.76	59.34	63.69
50	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
60	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
70	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4
80	53.34	57.15	60.39	64.28	96.58	101.9	106.6	112.3

Table ?? - Standard Normal Distribution

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw$$

$$\Phi(z) = 1 - \Phi(-z)$$

z = row + column

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1.0	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977
2.0	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982
2.1	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986
2.2	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989
2.3	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992
2.4	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994
2.5	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995
2.6	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
2.7	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999
3.0	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999

Quantiles

$P(Z > z_\alpha) = \alpha$

α	0.400	0.300	0.200	0.100	0.050	0.025	0.020	0.010	0.005	0.001
z_α	0.253	0.524	0.842	1.282	1.645	1.960	2.054	2.326	2.576	3.090
$z_{\alpha/2}$	0.842	1.036	1.282	1.645	1.960	2.240	2.326	2.576	2.807	3.291

Table ?? - Student's t Distribution

$$F(t) = P(T \le t) = \int_{-\infty}^{t} \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi r}\,\Gamma\left(\frac{r}{2}\right)\left(1+\frac{w^2}{r}\right)^{\frac{r+1}{2}}} dw$$

$$P(T \le -t) = 1 - P(T \le t)$$

	$P(T \le t)$						
	0.60	0.75	0.90	0.95	0.975	0.99	0.995
k	$t_{0.40}(k)$	$t_{0.25}(k)$	$t_{0.10}(k)$	$t_{0.05}(k)$	$t_{0.025}(k)$	$t_{0.01}(k)$	$t_{0.005}(k)$
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576

Part IV

TODO

- entropy
- probability generating function
- Fisher Information
- notation
- German Tank Problem
- continuity correction
- beta distribution
- irwin-hall distribution
- inverse transform sampling
- coefficient of variation
- index of dispersion
- convolution of probability distributions
- precision
- R functions after distribution tables?
- Power of a test

INDEX