

# STATS200 class notes

Erich Trieschman

2021 Fall quarter

## Contents

<b>1</b>	<b>Review: Combinatorics and probability</b>	<b>2</b>
1.1	Calculus cheat sheet	2
1.1.1	Logs	2
1.1.2	Derivatives	2
1.1.3	Integrals	3
1.1.4	Infinite series and sums	3
1.2	Events and sets	3
1.3	Probability	3
1.3.1	Conditional probability	4
1.3.2	Independence	4
<b>2</b>	<b>Random variables and common distribution functions</b>	<b>4</b>
2.1	Discrete distribution functions	4
2.1.1	Bernoulli	4
2.1.2	Binomial distribution	5
2.1.3	Geometric distribution	5
2.1.4	Negative binomial	5
2.1.5	Poisson distribution	5
2.2	Continuous distribution functions	5
2.2.1	Uniform distribution	5
2.2.2	Normal distribution	6
2.2.3	Exponential distribution	6
2.2.4	Gamma distribution	6
2.2.5	Cauchy distribution	6
2.2.6	Beta distribution	6
<b>3</b>	<b>Joint, marginal, and conditional distributions</b>	<b>7</b>
3.1	Joint distributions	7
3.1.1	Distribution of sums of independent random variables	7
3.1.2	Expectation of joint distributions	7
3.2	Marginal distributions	8
3.3	Conditional distributions	8
<b>4</b>	<b>Expected variables</b>	<b>8</b>
4.1	Expected value	8
4.2	Variance	9
4.3	Covariance	9
4.4	Correlation	10
4.5	Key theorems	10
4.5.1	Iterated expectation	10
4.5.2	Variance decomposition	10
4.5.3	Cauchy-Schwartz inequality	10
4.5.4	Jensen inequality	10
4.5.5	Markov inequality	10

4.5.6	Chebyshev inequality	11
4.6	Moment generating function	11
4.6.1	Common MGF derivations	11
<b>5</b>	<b>Convergence and limit theorems</b>	<b>11</b>
5.1	Convergence in probability	11
5.2	Convergence in $L_p$	11
5.3	Convergence in distribution	12
5.3.1	Convergence in probability $\implies$ convergence in distribution	12
5.3.2	Slutsky's theorem	12
5.3.3	Student's t distribution (example use case of Slutsky)	12
5.4	Law of large numbers	12
5.5	Central limit theorem	13
5.6	Delta method	13
<b>6</b>	<b>Estimation</b>	<b>14</b>
6.1	Mean Squared Error	14
6.2	Method of Moments estimator	14
6.3	Maximum likelihood estimator	15
6.4	Fisher Information	15
6.4.1	Properties of Fischer Information	16
6.4.2	The "Big" theorem: Asymptotic distribution using Fischer Information	16
6.5	Bayes estimator	17
6.6	Key theorems	17
6.7	Consistency	17
6.8	Efficiency	17
6.9	Sufficiency	17
<b>7</b>	<b>Hypothesis testing</b>	<b>17</b>
7.1	Likelihood ratio	17
7.2	Neyman-Pearson lemma	17
7.3	Uniformly Most Powerful tests	17
7.4	Confidence intervals	17
<b>8</b>	<b>Analysis of categorical data</b>	<b>17</b>
8.1	Chi-Square Test	17
8.2	Fisher's Exact Test	17

# 1 Review: Combinatorics and probability

## 1.1 Calculus cheat sheet

### 1.1.1 Logs

- $\log_b(M * N) = \log_b M + \log_b N$
- $\log_b(\frac{M}{N}) = \log_b M - \log_b N$
- $\log_b(M^k) = k \log_b M$
- $e^n e^m = e^{n+m}$

### 1.1.2 Derivatives

- $(x^n)' = nx^{n-1}$
- $(e^x)' = e^x$
- $(e^{u(x)})' = u'(x)e^x$

- $(\log_e(x))' = (\ln x)' = \frac{1}{x}$
- $(f(g(x)))' = f'(g(x))g'(x)$

### 1.1.3 Integrals

- TODO: Integration by parts

### 1.1.4 Infinite series and sums

- $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$
- $\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots = \sum_{n=0}^{\infty} x^n$  for  $|x| < 1$
- $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}$
- $(1 + \frac{a}{n})^n \rightarrow e^a$

## 1.2 Events and sets

Set operations follow commutative, associative, and distributive laws:

- Commutative:  $E \cup F = F \cup E$  and  $E \cap F = F \cap E$  (also written  $EF = FE$ )
- Associative:  $(E \cup F) \cup G = E \cup (F \cup G)$  and  $(E \cap F) \cap G = E \cap (F \cap G)$
- Distributive:  $(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$  and  $E \cap (F \cup G) = (E \cap F) \cup (E \cap G)$

**DeMorgan's Laws** relate the complement of a union to the intersection of complements:

- $(\cup_{i=1}^n E_i)^c = \cap_{i=1}^n E_i^c$
- $(\cap_{i=1}^n E_i)^c = \cup_{i=1}^n E_i^c$

## 1.3 Probability

A **probability space** is defined by a triple of objects  $(S, \mathcal{E}, P)$ :

- $S$ : Sample space
- $\mathcal{E}$ : Set of possible events within the sample space. Set of events are assumed to be  $\theta$ -field (below)
- $P$ : Probability for each event

A  **$\theta$ -field** is a collection of subsets  $\mathcal{E} \subset S$  that satisfy

- $\emptyset \in \mathcal{E}$
- $E \in \mathcal{E} \Rightarrow E^C \in \mathcal{E}$
- $E_i \in \mathcal{E}$  for  $1, 2, \dots \Rightarrow \cup_{i=1}^{\infty} E_i \in \mathcal{E}$

**Basic probability properties**

- $P(A^C) = 1 - P(A)$
- $P(\emptyset) = 0$
- $A \subset B \Rightarrow P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

The **law of total probability** relates marginal probabilities to conditional probabilities. For a partition,  $E_1, E_2, \dots$  of set,  $S$ , where a partition implies i)  $E_i, E_j$  are pairwise disjoint and ii)  $\cup_{i=1}^{\infty} E_i = S$ , then

$$P(A) = \sum_{i=1}^{\infty} P(A \cap E_i) = \sum_{i=1}^{\infty} P(A | E_i)P(E_i)$$

The **continuity of probability measures** state

- (i)  $E_1 \subset E_2 \subset \dots$  Let  $E_{\infty} = \cup_i E_i$ , then  $P(E_n) \rightarrow P(E_{\infty})$  as  $n \rightarrow \infty$
- (ii)  $E_1 \supset E_2 \supset \dots$  Let  $E_{\infty} = \cap_i E_i$ , then  $P(E_n) \rightarrow P(E_{\infty})$  as  $n \rightarrow \infty$

### 1.3.1 Conditional probability

The conditional probability is the probability of one event occurring, given the other event occurring. A reframing of conditional probability (see formula below) is the probability of both events occurring, divided by the marginal probability of one of the events occurring.

$$p_{X|Y}(x|y) = \frac{p(x, y)}{p_y(y)}$$

**Bayes Theorem** leverages conditional probabilities of measured events to glean conditional probabilities of un-measured events:

$$P(E_i | B) = \frac{P(B | E_i)P(E_i)}{\sum_{j=1}^{\infty} P(B | E_j)P(E_j)} = \frac{P(B | E_i)P(E_i)}{P(B)}$$

Where  $E_1, E_2, \dots$  form a partition of the sample space.

### 1.3.2 Independence

Events  $A$  and  $B$  are independent if  $P(A \cap B) = P(A)P(B)$

It is possible for events to be pairwise independent, but not mutually independent. For example, toss a pair of dice and let  $D_1$  be the number for die 1 and  $D_2$  be the number for die 2. Define  $E_i = \{D_i \leq 2\}$ . And define  $E_3 = \{3 \leq \max(D_1, D_2) \leq 4\}$ . These events are pairwise independent, but  $P(E_1 \cap E_2 \cap E_3) = 0$ , so they are not mutually independent.

## 2 Random variables and common distribution functions

**Random variables** are functions connecting a sample space to real numbers. They are formally defined as

$$\{\omega \in S : X(\omega) \leq t\} \in \mathcal{E}$$

For example, if coin tosses produce a sample space of {Heads, Tails}, a random variable can be the number of heads.

### 2.1 Discrete distribution functions

#### 2.1.1 Bernoulli

**Probability mass function** (*Bernouli*( $p$ )): Random variable  $X$  takes the value 1 with probability  $p$  and the value 0 with probability  $1 - p$

$$p(x) = p^x(1 - p)^{1-x}, x \in \{0, 1\}$$

**Expected value:**  $p$

**Variance:**  $p(1 - p)$

### 2.1.2 Binomial distribution

**Probability mass function ( $Bin(n, p)$ ):** For random variable  $X$ , the number of successes in  $n$  trials, the probability of observing  $j$  successes where each success has probability  $p$  is

$$P(X = j) = \binom{n}{j} p^j (1 - p)^{n-j}$$

**Expected value:**  $np$

**Variance:**  $np(1 - p)$

### 2.1.3 Geometric distribution

**Probability mass function ( $Geom(p)$ ):** For random variable  $X$ , the number of trials until the first success (included) with probability  $p$  is

$$P(X = j) = (1 - p)^{j-1} p$$

**Expected value:**  $\frac{1}{p}$

**Variance:**  $\frac{1-p}{p^2}$

### 2.1.4 Negative binomial

**Probability mass function ( $NB(r, p)$ ):** For random variable  $X$ , the number of successes,  $k$  before a specified number of failures,  $r$ , with probability of success  $p$  is

$$P(X = k) = \binom{k+r-1}{k} (1-p)^r p^k$$

**Expected value:**  $\frac{pr}{1-p}$

**Variance:**  $\frac{pr}{(1-p)^2}$

### 2.1.5 Poisson distribution

**Probability mass function ( $Pois(\lambda)$ ):** For random variable,  $X$ , the number of events,  $k$ , occurring in a fixed interval of time or space if these events occur with a known constant mean rate,  $\lambda$

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

**Expected value:**  $\lambda$

**Variance:**  $\lambda$

## 2.2 Continuous distribution functions

### 2.2.1 Uniform distribution

**$Unif(a, b)$ :** The distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds.[1] The bounds are defined by the parameters,  $a$  and  $b$ , which are the minimum and maximum values

$$pdf : f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$
$$cdf : F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] ; \\ 1 & \text{for } x > b \end{cases}$$

**Expected value:**  $\frac{1}{2}(a + b)$

**Variance:**  $\frac{1}{12}(b - a)^2$

### 2.2.2 Normal distribution

$N(\mu, \sigma)$

$$pdf : f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$cdf : F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

**Expected value:**  $\mu$

**Variance:**  $\sigma^2$

### 2.2.3 Exponential distribution

$Exp(\lambda)$ : the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate. It is a particular case of the gamma distribution.

$$pdf : f(x) = \lambda e^{-\lambda x}$$

$$cdf : F(x) = 1 - e^{-\lambda x}$$

**Expected value:**  $\frac{1}{\lambda}$

**Variance:**  $\frac{1}{\lambda^2}$

### 2.2.4 Gamma distribution

$Gamma(\alpha, \lambda)$ : a two-parameter family of continuous probability distributions.

$$pdf : f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \text{ where } \Gamma(\alpha) = (\alpha-1)! \text{ for any positive integer, } \alpha$$

$$cdf : F(x) = \frac{1}{\Gamma(\alpha)} \gamma(\alpha, \lambda x), \text{ where } \gamma(\alpha, x) = \int_0^x t^{\alpha-1} e^{-t} dt$$

**Expected value:**  $\frac{\alpha}{\lambda}$

**Variance:**  $\frac{\alpha}{\lambda^2}$

### 2.2.5 Cauchy distribution

$Cauchy(t, s)$ : The Cauchy distribution is often used in statistics as the canonical example of a "pathological" distribution since both its expected value and its variance are undefined

$$pdf : f(x) = \frac{1}{s\pi(1 + (x-t)/s)^2}, \text{ where } s \text{ is the scale parameter and } t \text{ is the location parameter}$$

$$cdf : \frac{1}{\pi} \arctan\left(\frac{x-t}{s}\right) + \frac{1}{2}$$

**Expected value:**  $DNE$

**Variance:**  $DNE$

### 2.2.6 Beta distribution

$Beta(\alpha, \beta)$ : a family of continuous probability distributions defined on the interval  $[0, 1]$  parameterized by two positive shape parameters that appear as exponents of the random variable and control the shape of the distribution.

$$pdf : f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \text{ where } x \in [0, 1], \text{ and } \Gamma(k) = (k-1)! \text{ for any positive integer } k$$

**Expected value:**  $\frac{\alpha}{\alpha+\beta}$

**Variance:**  $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

### 3 Joint, marginal, and conditional distributions

#### 3.1 Joint distributions

The cumulative density function (cdf) and probability mass function (pmf) satisfy respectively

$$\begin{aligned}\text{cdf: } F_{X_1, \dots, X_n}(x_1, \dots, x_n) &= P(X_1 \leq x_1, \dots, X_n \leq x_n) \\ \text{pmf: } f_{X_1, \dots, X_n}(x_1, \dots, x_n) &= P(X_1 = x_1, \dots, X_n = x_n)\end{aligned}$$

The joint density function  $f$  then satisfies, for  $E \subset \mathbb{R}^n$ ,

$$P((X_1, \dots, X_n) \in E) = \int \cdots \int_E f_{X_1, \dots, X_n} dx_1 \dots dx_n$$

When random variables are independent, the joint cdf and pmf satisfy respectively

$$\begin{aligned}\text{cdf: } P(X_1 \leq x_1, \dots, X_n \leq x_n) &= P(X_1 \leq x_1) \dots P(X_n \leq x_n) = \prod_{i=1}^n P(X_i \leq x_i) \\ \text{pmf: } P(X_1 = x_1, \dots, X_n = x_n) &= P(X_1 = x_1) \dots P(X_n = x_n) = \prod_{i=1}^n P(X_i = x_i)\end{aligned}$$

##### 3.1.1 Distribution of sums of independent random variables

The following combination of marginal distributions is called a **convolution**.

If  $X$  and  $Y$  have densities, the cdf of  $X + Y$  is

$$\begin{aligned}F_{X+Y}(t) &= P(X + Y \leq t) \\ &= P(X \leq t - y) \\ &= \int_{-\infty}^{\infty} P(X \leq t - y \mid Y = y) f_Y(y) dy, \text{ to get marginal distribution} \\ &= \int_{-\infty}^{\infty} F_X(t - y) f_Y(y) dy, \text{ since } X, Y \text{ independent}\end{aligned}$$

Likewise, the density of the sum is

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(t - y) f_Y(y) dy$$

And similarly for discrete random variables

$$p_{X+Y}(t) = \sum_{x=-\infty}^{\infty} p_Y(t - x) p_X(x)$$

##### 3.1.2 Expectation of joint distributions

For  $X, Y$  joint distribution,  $f_{X,Y}(x, y)$ , or probability mass function,  $p(x, y)$

$$\begin{aligned}\text{pmf: } E[g(X, Y)] &= \sum_s g(X(s), Y(s)) p(s) \\ &= \sum_x \sum_y g(x, y) \sum_{s: X(s)=x, Y(s)=y} p(s) \\ &= \sum_x \sum_y g(x, y) p(x, y)\end{aligned}$$

$$\text{pdf: } E[g(X, Y)] = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} g(x, y) f(x, y) dx dy$$

### 3.2 Marginal distributions

Marginal density functions or marginal probability mass functions are obtained by integrating or summing out the other variables

$$pmf : f_Y(y) = \sum_x yP(Y = y | x)pdf : f_Y(y) = \int_a^b f(x, y)dx, \text{ where } x \in [a, b]$$

### 3.3 Conditional distributions

Reminder:

$$p_{X|Y}(x|y) = \frac{p(x, y)}{p_Y(y)} \text{ and } f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

We can use conditional probabilities to restate the **law of total probability**:

$$P(E) = \int_{-\infty}^{\infty} P(E | X = x)f(x)dx$$

## 4 Expected variables

### 4.1 Expected value

$$E(X) = \sum_x xP(X = x)$$

Which can also be written as

$$E(X) = \sum_{s \in S} X(s)p(s), \text{ where } p(s) \text{ is the probability that element } s \in S \text{ occurs:}$$

Proof:

$$\begin{aligned} E(X) &= \sum_i x_i P(X = x_i), \text{ for } E_i = \{X = x_i\} = \{s \in S : X(s) = x_i\} \\ &= \sum_i x_i \sum_{s \in E_i} p(s) = \sum_i \sum_{s \in E_i} x_i p(s) \\ &= \sum_i \sum_{s \in E_i} X(s)p(s) = \sum_{s \in S} x_i p(s) \end{aligned}$$

This equation structure helps proof several properties of the expected value:

- $E(g(X)) = \sum_i g(x_i)p_X(x_i)$ , assuming  $g(x_i) = y_i$

Proof:

$$\begin{aligned} \sum_i g(x_i)p_X(x_i) &= \sum_j \sum_{i:g(x_i)=y_j} g(x_i)p_X(x_i) = \sum_j \sum_{i:g(x_i)=y_j} y_j p_X(x_i) \\ &= \sum_j y_j \sum_{i:g(x_i)=y_j} p_X(x_i) = \sum_j y_j P(g(X) = y_j) \\ &= E(g(X)) \end{aligned}$$

- $E(aX + b) = aE(X) + b$

$$E(aX + b) = \sum_{s \in S} (aX(s) + b)p(s) = a \sum_{s \in S} X(s)p(s) + \sum_{s \in S} bp(s) = aE(X) + b$$

- $E(X + Y) = E(X) + E(Y)$

$$E(X + Y) = \sum_{s \in S} (X(s) + Y(s))p(s) = \sum_{s \in S} X(s)p(s) + \sum_{s \in S} Y(s)p(s) = E(X) + E(Y)$$



## 4.2 Variance

$$\begin{aligned} \text{Var}(X) &= E((X - E(X)))^2 = \sigma^2 \\ SD &= \sqrt{\text{Var}(X)} = \sqrt{\sigma^2} = \sigma \end{aligned}$$

Several properties of variance follow from linearity of expectation:

$$(i) \text{Var}(X) = E(X^2) - \mu^2$$

$$\text{Var}(X) = E((X - \mu)^2) = E(X^2 - 2X\mu + \mu^2) = E(X^2 - 2\mu X + \mu^2)$$

$$\text{Var}(X) = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2$$

$$(ii) \text{Var}(aX + b) = a^2 \text{Var}(X)$$

$$\text{Var}(aX + b) = E((aX + b)^2) - E(aX + b)^2 = E(a^2X^2 + 2abX + b^2) - (aE(X) + b)^2$$

$$\text{Var}(aX + b) = a^2E(X^2) + 2abE(X) + b^2 - a^2E(X)^2 - 2abE(X) - b^2 = a^2E(X^2) - a^2E(X)^2 = a^2(E(X^2) - E(X)^2)$$

$$(iii) \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \text{ for } X, Y \text{ independent}$$

$$\text{Var}(X + Y) = E((X + Y)^2) - E(X + Y)^2 = E(X^2) + 2E(XY) + E(Y^2) - E(X^2) - 2E(X)E(Y) - E(Y)^2$$

$$\text{Var}(X + Y) = E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2, \text{ since } E(XY) = 0 \text{ (by independence) and } E(X) = E(Y) = 0 \text{ (WLOG)}$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

## 4.3 Covariance

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$$

Several properties of covariance follow from linearity of expectation

$$(i) \text{Cov}(X, X) = \text{Var}(X) :$$

$$\text{Cov}(X, X) = E[(X - E(X))(X - E(X))] = E[(X - E(X))^2] = \text{Var}(X)$$

$$(ii) \text{Cov}(X, Y) = E(XY) - E(X)E(Y) :$$

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY - E(Y)X - E(X)Y + E(X)E(Y))$$

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) = E(XY) - E(X)E(Y)$$

$$(iii) \text{ if } X, Y \text{ independent, then } \text{Cov}(X, Y) = 0$$

$$(iv) \text{Cov}(aX, bY) = ab \text{Cov}(X, Y) :$$

$$\text{Cov}(aX, bY) = E(abXY) - E(aX)E(bY) = ab(E(XY) - E(X)E(Y)) = ab \text{Cov}(X, Y)$$

$$(v) \text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z) :$$

$$\text{Cov}(X, Y + Z) = E(X(Y + Z)) - E(X)E(Y + Z)$$

$$\text{Cov}(X, Y + Z) = E(XY) + E(XZ) - E(X)E(Y) - E(X)E(Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$$

$$(vi) \text{Cov}(U, V) = \sum_i \sum_j b_i d_j \text{Cov}(X_i, Y_j), \text{ with } U = a + \sum_i b_i X_i \text{ and } V = c + \sum_j d_j Y_j :$$

$$(vii) \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) :$$

$$\text{Var}(X + Y) = \text{Cov}(X + Y, X + Y) = \text{Cov}(U, V), \text{ for } U = V = X + Y$$

$$\text{Var}(X + Y) = \text{Cov}(U, V) = \text{Cov}(X, X) + \text{Cov}(X, Y) + \text{Cov}(Y, Y) + \text{Cov}(Y, X), \text{ using vi}$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

## 4.4 Correlation

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

## 4.5 Key theorems

### 4.5.1 Iterated expectation

**Law of iterated expectation:**  $E(E(Y | X)) = E(Y)$

**Proof:**

$$\begin{aligned} E(Y | X) &= \sum_y y \frac{f_{X,Y}(X, y)}{f_X(X)} \\ E(E(Y | X)) &= \sum_x \sum_y \left( y \frac{f_{X,Y}(x, y)}{f_X(x)} \right) f_X(x) = \sum_x \sum_y y f_{X,Y}(x, y) = \sum_y y f_Y(y) = E(Y) \end{aligned}$$

### 4.5.2 Variance decomposition

**Variance decomposition formula:**  $Var(Y) = E(Var(Y | X)) + Var(E(Y | X))$

### 4.5.3 Cauchy-Schwartz inequality

**Cauchy-Schwartz inequality:**  $E(UV)^2 \leq E(U^2)E(V^2)$ , with equality if  $P(cU = V) = 1$  for some constant,  $c$

**Proof:**

$$\begin{aligned} \text{let } h(t) &= E((tU - V)^2) \geq 0 \\ h(t) &= t^2 E(U^2) - 2t E(UV) + E(V^2), \text{ a quadratic equation} \\ h(t) \geq 0 &\Rightarrow \text{discriminant} \leq 0 \\ &\Rightarrow 4E(UV)^2 - 4E(U^2)E(V^2) \leq 0 \\ &\Rightarrow E(UV)^2 \leq E(U^2)E(V^2) \end{aligned}$$

### 4.5.4 Jensen inequality

**Jensen inequality:**  $E(g(x)) \geq g(E(x))$  for  $g(x)$  convex

**Proof:**

Let  $E(X) = \mu$ , and  $L(X)$  a line s.t.  $L(\mu) = g(E(x))$  :

$$g(X) \geq L(X) \text{ for all } X$$

$$E(g(X)) \geq E(L(X)) = L(E(X)) = g(E(X))$$

### 4.5.5 Markov inequality

**Markov inequality:** For  $X \geq 0$ ,  $P(X \geq t) \leq \frac{E(X)}{t} \quad \forall t > 0$

**Proof:**

$$\text{Let } y = \begin{cases} 1 & X \geq t \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Then } tY \leq X \text{ since } \begin{cases} X \geq t & t * 1 \leq X \\ X < t & t * 0 < X \end{cases}$$

$$tY \leq X \implies E(tY) \leq E(X) \implies tP(X \geq t) \leq E(X) \implies P(X \geq t) \leq \frac{E(X)}{t}$$

#### 4.5.6 Chebyshev inequality

**Chebyshev inequality:**  $P(|X - E(X)| \geq t) \leq \frac{Var(X)}{t^2} \quad \forall t > 0$

**Proof:**

$$\begin{aligned}
P(|X - E(X)| \geq t) &= P((X - E(X))^2 \geq t^2) \\
P((X - E(X))^2 \geq t^2) &\leq \frac{E((X - E(X))^2)}{t^2}, \text{ by Markov inequality} \\
P((X - E(X))^2 \geq t^2) &\leq \frac{Var(X)}{t^2}
\end{aligned}$$

### 4.6 Moment generating function

The moment generating function of a random variable  $X$  is defined as

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{n=0}^{\infty} \frac{\mathbb{E}[X^n]}{n!} t^n \leftarrow \text{power series}$$

Notice its called a moment generating function because each derivative of this function can generate a new moment of  $X$  at  $t = 0$ :

$$M_X^{(n)}(0) = \mathbb{E}[X^n]$$

#### 4.6.1 Common MGF derivations

- $Y = a + bX \implies M_Y = e^{at} M_X(bt)$
- $Z = X + Y, X \perp Y \implies M_Z = M_Y M_X = E(e^{tX}) E(e^{tY})$

## 5 Convergence and limit theorems

### 5.1 Convergence in probability

A sequence of random variables,  $X_n$ , converges in probability,  $X_n \xrightarrow{p} X$  when

$$P(|X_n - X| > \epsilon) \longrightarrow 0 \text{ as } n \longrightarrow \infty$$

**Consistent estimator:** An estimator,  $T_n = T_n(X_1, \dots, X_n)$ , which converges in probability to  $g(\theta)$ , a function of the model parameter

**Additional properties** of convergence in probability

- if  $X_n \xrightarrow{p} X$  and  $a_n \xrightarrow{p} a$  then  $a_n X_n \xrightarrow{p} aX$
- if  $X_n \xrightarrow{p} X$  and  $A_n \xrightarrow{p} A$  then  $A_n X_n \xrightarrow{p} AX$
- if  $X_n \xrightarrow{p} X$ ,  $A_n \xrightarrow{p} A$ , and  $B_n \xrightarrow{p} B$  then  $A_n X_n + B_n \xrightarrow{p} AX + B$
- if  $X_n \xrightarrow{p} X$  and  $g$  a continuous function then  $g(X_n) \xrightarrow{p} g(X)$  (**continuous mapping theorem**)

### 5.2 Convergence in $L_p$

See [https://en.wikipedia.org/wiki/Lp\\_space](https://en.wikipedia.org/wiki/Lp_space) for more information (not much covered in class).

Convergence in  $L_p$  is stronger than convergence in probability. **Counter example** to convergence in probability  $\implies$  convergence in  $L_p$ :

$$\text{Let } X_n = \begin{cases} n & \frac{1}{n} \\ 0 & 1 - \frac{1}{n} \end{cases}$$

$$X_n \xrightarrow{p} 0 : P(|X_n - 0| \geq \epsilon) = P(X_n = n) = 1/n \longrightarrow 0 \text{ as } n \longrightarrow \infty$$

$$\text{but } E(X_n) = n \frac{1}{n} + 0(1 - \frac{1}{n}) = 1 \implies \text{no convergence in } L_p$$

### 5.3 Convergence in distribution

A sequence of random vectors,  $X_n$ , converges in distribution to a random vector,  $X_n \xrightarrow{d} X$  when

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x) \text{ at all continuity points in } F_X$$

- Convergence in distribution **does not** imply convergence in probability unless convergence in distribution is to a single point
- if  $X_n \xrightarrow{d} X$  and  $g$  a continuous function then  $g(X_n) \xrightarrow{d} g(X)$  (**continuous mapping theorem**)

#### 5.3.1 Convergence in probability $\implies$ convergence in distribution

Let  $X$  have cdf,  $F$ , with  $t$  a continuity point of  $F$

$$\begin{aligned} P(X_n \leq a) &\leq P(X \leq a + \epsilon) + P(|X_n - X| > \epsilon) \text{ by lemma} \\ P(X \leq a - \epsilon) - P(|X_n - X| > \epsilon) &\leq P(X_n \leq a) \leq P(X \leq a + \epsilon) + P(|X_n - X| > \epsilon) \\ F_X(a - \epsilon) &\leq \lim_{n \rightarrow \infty} P(X_n \leq a) \leq F_X(a + \epsilon), \text{ where } F_X(a) = P(X \leq a) \\ &\implies \lim_{n \rightarrow \infty} P(X_n \leq a) = P(X \leq a) \implies \{X_n\} \xrightarrow{d} X \end{aligned}$$

#### 5.3.2 Slutsky's theorem

$A_n X_n + B_n \xrightarrow{d} aX + b$  if

- $\{X_n\}$  sequence with  $X_n \xrightarrow{d} X$
- $\{A_n\}$  sequence with  $A_n \xrightarrow{d} A$
- $\{B_n\}$  sequence with  $B_n \xrightarrow{d} b$

#### 5.3.3 Student's t distribution (example use case of Slutsky)

$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}} \xrightarrow{d} N(0, 1)$ :

$$\begin{aligned} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}} &= \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \frac{\sigma}{\hat{\sigma}} \\ \text{And we know } \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} &\xrightarrow{d} N(0, 1) \\ \text{and } \frac{\sigma}{\hat{\sigma}} &\xrightarrow{p} 1 \text{ since } \hat{\sigma} \xrightarrow{p} \sigma \\ \text{So, by Slutsky's theorem } \frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}} &\xrightarrow{d} N(0, 1) * 1 \end{aligned}$$

**This RHS term is referred to as the t-statistic**, which follows a Student's t distribution with  $n - 1$  degrees of freedom. In practice, if the sample is reasonably sized, it won't make a difference using the Normal distribution instead of the Student's t distribution.

### 5.4 Law of large numbers

For  $X_1, X_2, \dots, X_n$  a sequence of i.i.d. random variables with  $E(X_i) = \mu$ ,  $Var(X_i) = \sigma^2$ ,  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ , then for any  $\epsilon > 0$

$$P(|\bar{X}_n - \mu| > \epsilon) \longrightarrow 0 \text{ as } n \rightarrow \infty$$

**Proof:**

First find  $\mathbb{E}(\bar{X}_n)$  and  $Var(\bar{X}_n)$

$$\mathbb{E}(\bar{X}_n) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) = \mu$$

$$Var(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) = \frac{\sigma^2}{n}, \text{ since } X_i \text{ independent}$$

The desired result now follows immediately from Chebyshev's inequality, which states

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{Var(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \rightarrow 0 \text{ as } n \rightarrow \infty$$

## 5.5 Central limit theorem

Most useful form of CLT, which can be used for approximate methods:

$$\sqrt{n} \frac{(\bar{X}_n - \mu)}{\sigma} \rightarrow N(0, 1)$$

$$\sqrt{n}(\bar{X}_n - \mu) \rightarrow N(0, \sigma^2)$$

**More formal definition and proof:** For  $X_1, X_2, \dots, X_n$  a sequence of i.i.d. random variables with  $E(X_i) = 0$ ,  $Var(X_i) = \sigma^2$ , and the common cumulative distribution function  $F$  and moment-generating function  $M$  defined in a neighborhood of zero. Then

$$\text{For } S_n = \sum_{i=1}^n X_i$$

$$\lim_{n \rightarrow \infty} P\left(\frac{S_n}{\sigma\sqrt{n}} \leq x\right) = \Phi(x)$$

**Proof:** Let  $Z_n = \frac{S_n}{\sigma\sqrt{n}}$ . We show the MGF of  $Z_n$  tends to the MGF of the standard normal distribution. Since  $S_n$  is a sum of independent random variables,

$$M_{S_n}(t) = [M(t)]^n \text{ and } M_{Z_n}(t) = [M(\frac{t}{\sigma\sqrt{n}})]^n$$

Reminder: Taylor series expansion of  $M(s) = M(0) + sM'(0) + \frac{1}{2}sM''(0) + \epsilon_s$

$$M(\frac{t}{\sigma\sqrt{n}}) = 1 + \frac{1}{2}\sigma^2(\frac{t}{\sigma\sqrt{n}})^2 + \epsilon_n \text{ with } E(X) = M'(0) = 0, Var(X) = M''(0) = \sigma^2$$

$$M_{Z_n}(t) = (1 + \frac{t^2}{2n} + \epsilon_n)^n$$

$$M_{Z_n}(t) \rightarrow e^{\frac{t^2}{2}} \text{ as } n \rightarrow \infty, \text{ by the infinite series convergence to } e^a$$

Since  $e^{\frac{t^2}{2}}$  is the MGF of the standard normal distribution, we have proven the central limit theorem.

## 5.6 Delta method

If  $g$  is a differentiable function at  $\mu$

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} N(0, g'(\mu)^2 \sigma^2)$$

## Proof

For general  $g$  and assuming  $E(\bar{X}_n) = \mu$  :

$$g(\bar{X}_n) \approx g(\mu) + g'(\mu)(\bar{X}_n - \mu) + \frac{1}{2}g''(\mu)(\bar{X}_n - \mu)^2 + \epsilon \text{ (Taylor approximation of } g(\mu))$$

$$g(\bar{X}_n) - g(\mu) \approx g'(\mu)(\bar{X}_n - \mu) + \epsilon$$

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) \approx g'(\mu)\sqrt{n}(\bar{X}_n - \mu) + \epsilon \text{ and we know}$$

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

$$g'(\mu)\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, g'(\mu)^2\sigma^2)$$

$$\text{So } \sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} N(0, g'(\mu)^2\sigma^2)$$

**Note:** if we find that  $g'(\mu) = 0$ , then repeat this process with the second derivative,  $g''(\mu)$ . The end result is a formula that converges in distribution to a scaling of a random variable,  $Z^2$  which follows a  $\chi_1^2$  distribution.

## 6 Estimation

The following section provides an overview of methods for estimating population parameters,  $\theta$ , using functions of the data ("estimators"),  $T(X_1, \dots, X_n)$

### 6.1 Mean Squared Error

The **Mean Squared Error (MSE)** can be used to evaluate our estimators.

$$\begin{aligned} MSE(T, \theta) &= E_\theta[(T - g(\theta))^2] \\ &= E_\theta(T^2) - 2g(\theta)E_\theta(T) + g(\theta)^2 \\ &= Var_\theta(T) + E_\theta(T)^2 - 2g(\theta)E_\theta(T) + g(\theta)^2 \\ &= Var_\theta(T) + (E_\theta(T) - g(\theta))^2 \\ &= Var_\theta(T) + Bias_\theta^2(T), \text{ where } Bias_\theta(T) = E_\theta(T) - g(\theta) \end{aligned}$$

### 6.2 Method of Moments estimator

To generate a method of moments estimator

- Calculate a moment using the moment generating function of the assumed distribution. Any moment,  $k$ , can be used, but lower moments will typically lead to an estimator distribution with lower variance

$$E(X^k) = g(\theta)$$

- Invert this expression to create an expression for the parameter(s) in terms of the moment

$$g^{-1}(E(X^k)) = \theta \implies f(E(X^k)) = \theta, \text{ where } f(x) = g^{-1}(x)$$

- Insert the sample moment into this expression, thus obtaining estimates of the parameters in terms of data

$$\hat{\theta} = f\left(\frac{1}{n} \sum X_i^k\right), \text{ by LNN } \frac{1}{n} \sum X_i^k \xrightarrow{p} E(X^k)$$

- Use the delta method to determine what the method of moments estimator converges to in distribution

$$\sqrt{n}\left(f\left(\frac{1}{n} \sum X_i^k\right) - \theta\right) \xrightarrow{d} N(0, f'(E(X_i^k))^2 Var(X_i^k)^2)$$

Methods of moment estimators are not uniquely determined, nor must they exist. The motivation for subsequent estimators is to help us pick the estimator with the smallest possible variance.

### 6.3 Maximum likelihood estimator

The **maximum likelihood estimator** constructs an estimator,  $\hat{\theta}_{MLE}$ , that maximizes the likelihood function with respect to  $\theta$ .

The **likelihood function**,  $L(\theta)$  is the joint density or probability mass function,  $f(X, \theta)$  evaluated at the data,  $\{X_i, \dots, X_n\}$ . Assuming the data is *i.i.d.*:

$$L(\theta) = \prod_{i=1}^n f(X_i, \theta)$$

The typical method to generate a maximum likelihood estimator

- Construct the likelihood function

$$L(\theta) = \prod_{i=1}^n f(X_i, \theta)$$

- Take the log of the likelihood function (usually leading to a function that is easier to derive)

$$\log(L(\theta)) = l(\theta) = \sum_{i=1}^n \log(f(X_i, \theta))$$

- Take the derivative of the log-likelihood function with respect to  $\theta$

$$\frac{d}{d\theta} l(\theta) = \sum_{i=1}^n \frac{d}{d\theta} \log(f(X_i, \theta))$$

- Find critical points of this function and determine that one is a maximum

$$\begin{aligned} 0 &= \sum_{i=1}^n \frac{d}{d\theta} \log(f(X_i, \hat{\theta})) \\ 0 &= \sum_{i=1}^n \frac{d^2}{d\theta^2} \log(f(X_i, \hat{\theta})), \text{ checking if } \hat{\theta} < 0 \end{aligned}$$

See next section on **Fischer Information** for guidance on the asymptotic distribution of the maximum likelihood estimator

### 6.4 Fisher Information

The **information** that data,  $X$ , contains about parameter,  $\theta$  is defined by

$$I(\theta) = E_{\theta} \left[ \left( \frac{d}{d\theta} \log(f(X, \theta)) \right)^2 \right]$$

- Fisher Information assumes **differentiability** and **existence of the second moment**
- $\frac{d}{d\theta} \log(f(X, \theta))$  is called the **score** function

### 6.4.1 Properties of Fischer Information

$$(i) E_{\theta} \left[ \left( \frac{d}{d\theta} \log(f(X, \theta)) \right) \right] = 0 :$$

$$E_{\theta} \left[ \left( \frac{d}{d\theta} \log(f(X, \theta)) \right) \right] = \int \frac{d}{d\theta} \log(f(x, \theta)) f(x, \theta) dx = \int \frac{f'(x, \theta)}{f(x, \theta)} f(x, \theta) dx = \int f'(x, \theta) dx$$

$$E_{\theta} \left[ \left( \frac{d}{d\theta} \log(f(X, \theta)) \right) \right] = \frac{d}{d\theta} \int f(x, \theta) dx = \frac{d}{d\theta} * 1 = 0$$

$$(ii) I(\theta) = Var \left( \frac{d}{d\theta} \log(f(X, \theta)) \right) :$$

$$Var \left( \frac{d}{d\theta} \log(f(X, \theta)) \right) = E_{\theta} \left[ \left( \frac{d}{d\theta} \log(f(X, \theta)) \right)^2 \right] - E_{\theta} \left[ \left( \frac{d}{d\theta} \log(f(X, \theta)) \right) \right]^2$$

$$Var \left( \frac{d}{d\theta} \log(f(X, \theta)) \right) = I(\theta) - 0^2 = I(\theta)$$

$$(iii) I(\theta) = -E_{\theta} \left[ \frac{d^2}{d\theta^2} \log(f(X, \theta)) \right]$$

$$\frac{d}{d\theta} \log(f(x, \theta)) = \frac{f'(x, \theta)}{f(x, \theta)} \implies \frac{d^2}{d\theta^2} \log(f(x, \theta)) = \frac{f(x, \theta)f''(x, \theta) - f'(x, \theta)^2}{f(x, \theta)^2}$$

$$E \left[ \frac{d^2}{d\theta^2} \log(f(x, \theta)) \right] = \int \frac{f(x, \theta)f''(x, \theta) - f'(x, \theta)^2}{f(x, \theta)^2} f(x, \theta) dx = \int f''(x, \theta) - I(\theta)$$

$$E \left[ \frac{d^2}{d\theta^2} \log(f(x, \theta)) \right] = -I(\theta), \text{ since } \int \frac{d^2}{d\theta^2} f(x, \theta) = \frac{d^2}{d\theta^2} * 1 = 0$$

$$(iv) I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta) \text{ for } X, Y \text{ independent}$$

**Corrolary:**  $I_n(\theta) = nI_1(\theta)$  for  $X_1, \dots, X_n$  i.i.d with  $I_1(\theta)$  the Information based on one data

**Note:** Information increases with larger sample!

$$(v) \text{ Cramer-Rau-Fisher Inequality: } Var(T(X)) \geq \frac{g'(\theta)^2}{I(\theta)} \text{ for } E(T(X)) = g(\theta)$$

$$Cov[T(X), \frac{d}{d\theta} \log(f(X, \theta))] = E[T(X) \frac{d}{d\theta} \log(f(X, \theta))], \text{ using property 1}$$

$$Cov[T(X), \frac{d}{d\theta} \log(f(X, \theta))] = \int T(x) f'(x, \theta) dx = \frac{d}{d\theta} \int T(x) f(x, \theta) dx = \frac{d}{d\theta} E(T(X)) = \frac{d}{d\theta} g(\theta) = g'(\theta)$$

$$g'(\theta)^2 \leq Var(T(X)) Var \left( \frac{d}{d\theta} \log(f(X, \theta)) \right) = Var(T(X)) I(\theta) \text{ by correlation inequality: } \rho^2 \leq 1$$

$$Var(T(X)) \geq \frac{g'(\theta)^2}{I(\theta)}$$

### 6.4.2 The "Big" theorem: Asymptotic distribution using Fischer Information

Under regularity assumptions, the maximum likelihood estimator (or any other reasonable estimator),  $\hat{\theta}$  of  $\theta$  satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N \left( 0, \frac{1}{I(\theta)} \right)$$



Sketch of proof:

$$L(\theta) = \prod_{i=1}^n f(X_i, \theta)$$

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^n \log(f(X_i, \theta))$$

MLE solves  $l'(\hat{\theta}) = 0$ , with  $l'(\theta) \approx l'(\theta_0) + (\hat{\theta} - \theta_0)l''(\theta_0)$

(full proof requires showing the error in this approx. is small)

$$0 = l'(\theta_0) + (\hat{\theta} - \theta_0)l''(\theta_0) \implies \hat{\theta} - \theta_0 = \frac{l'(\theta_0)}{l''(\theta_0)}$$

$$\implies \sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n} \frac{l'(\theta_0)}{l''(\theta_0)} = \frac{l'(\theta_0)}{\sqrt{n}} \div \frac{l''(\theta_0)}{n}$$

$$\frac{l''(\theta_0)}{n} = \frac{\sum \frac{d^2}{d\theta^2} \log(f(X, \theta))}{n} \xrightarrow{p} -E_{\theta} \left[ \frac{d^2}{d\theta^2} \log(f(X, \theta)) \right] = I(\theta)$$

$$\frac{l'(\theta_0)}{\sqrt{n}} = \frac{\sum \frac{d}{d\theta} \log(f(X, \theta))}{\sqrt{n}} \xrightarrow{d} N(0, I(\theta))$$

$$\frac{l'(\theta_0)}{\sqrt{n}} \div \frac{l''(\theta_0)}{n} \xrightarrow{d} N\left(0, \frac{I(\theta)}{I(\theta)^2}\right) = N\left(0, \frac{1}{I(\theta)}\right), \text{ by Slutsky's theorem}$$

## 6.5 Bayes estimator

## 6.6 Key theorems

## 6.7 Consistency

## 6.8 Efficiency

## 6.9 Sufficiency

# 7 Hypothesis testing

## 7.1 Likelihood ratio

## 7.2 Neyman-Pearson lemma

## 7.3 Uniformly Most Powerful tests

## 7.4 Confidence intervals

# 8 Analysis of categorical data

## 8.1 Chi-Square Test

## 8.2 Fisher's Exact Test