Project C-B Project

Charlie Windolf

Contents

1 Probability Theory **3**
- 1.1 Set theory . . . 3
- 1.2 Basics of probability theory . . . 4
- 1.3 Conditional probability and independence . . . 5
- 1.4 Random variables . . . 6
- 1.5 Distribution functions . . . 6
- 1.6 Important examples . . . 7

2 Transformations and Expectations **8**
- 2.1 Distributions of functions of a random variable . . . 8
- 2.2 Expectation . . . 9
- 2.3 Moment generating functions . . . 9
- 2.4 Differentiation under the integral sign . . . 10

3 Common Families of Distributions **12**
- 3.1 Discrete distributions . . . 12
- 3.2 Continuous distributions . . . 13
- 3.3 Exponential families . . . 14
- 3.4 Location and scale families . . . 15
- 3.5 Probability inequalities . . . 15
- 3.6 Identities . . . 15

4 Multiple Random Variables **16**
- 4.1 Joint and marginal distributions . . . 16
- 4.2 Conditional distributions and independence . . . 16
- 4.3 Bivariate transformations . . . 16
- 4.4 Hierarchical models and mixture distributions . . . 17
- 4.5 Covariance and correlation . . . 18
- 4.6 Multivariate distributions . . . 18
- 4.7 Inequalities . . . 19

5 Properties of a Random Sample **20**
- 5.1 Basic concepts of random samples . . . 20
- 5.2 Sampling from the normal distribution . . . 21
- 5.3 Order statistics . . . 22
- 5.4 Convergence concepts . . . 22

5.5 Generating a random sample 23

Chapter 1

Probability Theory

Fix a sample space S and for all $A \subset S$ define $A^C = S \setminus A$.

1.1 Set theory

Theorem 1.1.1 (Set arithmetic)**.** *Set intersection $\cup$ and union $\cap$ are commutative, associative, and distributive, i.e.*

- (*Commutativity*)
$$A \cup B = B \cup A,$$
$$A \cap B = B \cap A.$$
- (*Associativity*)
$$A \cup (B \cup C) = (A \cup B) \cup C,$$
$$A \cap (B \cap C) = (A \cap B) \cap C,$$
- (*Distributivity*)
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$
- (*De Morgan*)
$$(A \cup B)^C = A^C \cap B^C,$$
$$(A \cap B)^C = A^C \cup B^C.$$

De Morgan's laws extend to arbitrary unions and intersections. Actually, by definition of set membership, all of the above also extend to arbitrary unions and intersections.

Definition 1.1.1 (Disjoint events, partition)**.** $A_1, A_2, \ldots$ are *pairwise disjoint* if for all $i \neq j$, $A_i \cap A_j = \varnothing$. If $\{A_i\}_{i \in I}$ are pairwise disjoint and $\bigcup_{i \in I} A_i = S$, then $\{A_i\}$ are said to form a *partition* of S.

1.2 Basics of probability theory

Definition 1.2.1 (σ-algebra)**.** A collection $\mathcal{B}$ of subsets of S is called a *σ-algebra* if

a. $\varnothing \in \mathcal{B}$,

b. Closed under complement: $A \in \mathcal{B} \implies A^C \in \mathcal{B}$,

c. Closed under countable union: $A_1, A_2, \cdots \in \mathcal{B} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$.

Definition 1.2.2 (Borel algebra)**.** The Borel algebra is the smallest σ-algebra containing all open sets. We'll use $\mathcal{B}_{\mathbb{R}}$ to denote the Borel algebra on the reals.

Note that (a.,b.) $\implies S = \varnothing^C \in \mathcal{B}$, and by De Morgan and (c.), $\mathcal{B}$ is also closed under countable intersections. A set S together with a σ-algebra $\mathcal{B}$ is called a *measurable space*.

Definition 1.2.3 (Probability measure)**.** Given a measurable space $(S, \mathcal{B})$, a *probability measure* is a set function $P : \mathcal{B} \to \mathbb{R}$ such that

1. $P(A) \geq 0$ for all A,
2. $P(S) = 1$,
3. For $A_1, A_2, \ldots$ pairwise disjoint, $P(\bigcup_1^{\infty} A_i) = \sum_1^{\infty} P(A_i)$.

Axiom of Continuity. If $A_1 \supset A_2 \supset \ldots$ and $\bigcap_n A_n = \varnothing$, then $P(A_n) \to 0$.

As shown in an exercise, continuity combined with the finite union property imply the countable union property.

Proposition 1.2.1 (Basic properties of probability measures)**.** Let $(S, \mathcal{B}, P)$ be a probability space and A a measurable set. Then,

a. $P(\varnothing) = 0$,

b. $P(A) \leq 1$,

c. $P(A^C) = 1 - P(A)$,

d. $P(B \cap A^C) = P(B) - P(B \cap A)$,

e. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, which weakens to $P(A \cap B) \geq P(A) + P(B) - 1$ (Bonferroni),

f. $A \subset B \implies P(A) \leq P(B)$.

Further,

g. $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$ for any partition $\{C_i\}$.

h. $P(\bigcup_1^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ for any $\{A_i\} \subset \mathcal{B}$ (Boole).

Note (On counting).

- Recall that the number of independent choices from m collections of sizes $n_1, \ldots, n_m$ is $n_1 \times \cdots \times n_m$.

- How many ways are there to sample k objects from a collection of size n? Well, it depends on the sampling method:
  - *Ordered, with replacement*: n^k.
  - *Ordered, without replacement*:

$$\frac{n!}{(n-k)!} = n \times (n-1) \times \cdots \times (n-k+1).$$

  - *Unordered, without replacement*:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n}{k} \times \frac{n-1}{k-1} \times \cdots \times \frac{n-k+1}{1}.$$

  - *Unordered, with replacement*, aka *stars and bars*: since we're putting k balls into n bins, we are equivalently arranging $n-1$ bars and k stars. So, there are $k+n-1$ total slots, and we want to choose k of them for the stars, putting us in the previous case:

$$\frac{(k+n-1)!}{k!(n-1)!} = \binom{k+n-1}{k}.$$

1.3 Conditional probability and independence

Definition 1.3.1 (Conditional probability)**.** Let A, B be measurable with $P(B) > 0$. Then the *conditional probability* $P(A \mid B)$ is given by

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

Theorem 1.3.1 (Bayes' rule)**.** *For B measurable and $A_1, A_2, \ldots$ a partition of the sample space, we have*

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B \mid A_j)P(A_j)},$$

and in particular for any measurable A,

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

To obtain the latter, take $A = \bigcup_{i=1}^{\infty} A_i$.

Proposition 1.3.1 (Chain rule)**.** For any events $A_1, \ldots, A_n$,

$$\begin{aligned} P(A_n \cap \cdots \cap A_1) &= P(A_n \mid A_{n-1} \cap \cdots \cap A_1) \times \cdots \times P(A_2 \mid A_1)P(A_1) \\ &= \prod_{k=1}^{n} P\Big(A_k \mid \bigcap_{j=1}^{k-1} A_j\Big). \end{aligned}$$

This follows by straight induction from $P(A \cap B) = P(B \mid A)P(A)$.

Definition 1.3.2 (Independence)**.** Two events A, B are said to be *independent*, written $A \perp\!\!\!\perp B$, if

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

Note: If $A \perp\!\!\!\perp B$, then also $A \perp\!\!\!\perp B^C$, $A^C \perp\!\!\!\perp B$, $A^C \perp\!\!\!\perp B^C$.

Definition 1.3.3 (Mutual independence)**.** A collection of events $\{A_i\}_{i \in I}$ is said to be mutually independent if for all $J \subset I$,

$$P\Big(\bigcap_{j \in J} A_j\Big) = \prod_{j \in J} P(A_j).$$

1.4 Random variables

Definition 1.4.1 (Measurable function)**.** Let $(S, \mathcal{A})$, $(T, \mathcal{B})$ be measurable spaces and $f : A \to B$. Then we say that f is measurable if $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$, i.e. preimages of measurable sets are measurable.

Note: The book ignores the measurability of random variables, but I wanted to write it down.

Definition 1.4.2 (Random variable)**.** A random variable is a measurable function $X : (S, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

Definition 1.4.3 (Pushforward, induced probability measure)**.** Let X be a random variable with range $\mathcal{X} = X(S)$. Then X induces a *pushforward* probability measure on $\mathcal{X}$, $P_X = P \circ X^{-1}$, i.e.

$$P_X(E) = P(X^{-1}(E)).$$

This will most often be written as $P(X \in E)$.

1.5 Distribution functions

Definition 1.5.1 (Cumulative distribution function)**.** The *cumulative distribution function* or *cdf* of a random variable X is

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

Note: In the book, they have (to me) some problems with notation where $F_X(x) = P_X(X \leq x)$. This does not make sense to me, since $X \leq x$ lives in the sample space and is not a subset of $\mathbb{R}$ under any reasonable interpretation.

Definition 1.5.2 (cádlág)**.** Continuous from the right, with limits existing from the left.

Theorem 1.5.1 (Properties of the .)**.** *$F(x)$ is a cdf if and only if the following hold*

a. $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$,

b. $F(x)$ *is nondecreasing,*

c. $F(x)$ *is right continuous.*

Note that (b.) and (c.) imply cádlág. Actually that's not really important at this point?

Note: We say that a random variable with a continuous cdf is *continuous* and that one with a step function cdf is *discrete*.

Definition 1.5.3 (Identically distributed)**.** Random variables X, Y are *identically distributed* if $P_X = P_Y$, i.e. if for all $A \in \mathcal{B}_{\mathbb{R}}$, $P(X \in A) = P(Y \in A)$.

Definition 1.5.4 (pdf, pmf)**.** The definition they give is vague. But the point is these are positive functions that sum to 1 over the range of the random variable.

1.6 Important examples

- Binomial pmf: If $X \sim \text{Bin}(n, p)$, then

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

Chapter 2

Transformations and Expectations

2.1 Distributions of functions of a random variable

The first thing is, if $Y = g(x)$ for some $g : \mathcal{X} \to \mathcal{Y}$, we have

$$P(Y \in A) = P(g(X) \in A) = P(X \in g^{-1}(A)).$$

- For discrete random variables X, we can find the pmf f_Y of Y as

$$f_Y(y) = \sum_{x \in g^{-1}(y)} f_X(x).$$

- For continuous random variables, we have

$$F_Y(y) = \int_{\{x : g(x) \leq y\}} f_X(x)\, dx.$$

Theorem 2.1.1 (Monotone transformations)**.** *Let $\mathcal{X}$ the support of X and $\mathcal{Y}$ the image of $\mathcal{X}$ through g.*

- *If g is increasing on $\mathcal{X}$, $F_Y(y) = F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.*
- *If g is increasing, $F_Y(y) = 1 - F_X(g^{-1}(y))$.*

Example 2.1.1 (Log uniform is exponential). For X uniform on $[0, 1]$, $-\log X$ is exponentially distributed with cdf $1 - e^{-y}$.

Theorem 2.1.2 (Obtaining the pdf by the chain rule)**.** *Let $\mathcal{X}$ the support of X with pdf f_X, and $\mathcal{Y}$ the image of $\mathcal{X}$ through g. Say that f_X is continuous and that g^{-1} is continuously differentiable on $\mathcal{Y}$. Then*

$$f_Y(y) = f_X(g^{-1}(y)) \left| \tfrac{d}{dy} g^{-1}(y) \right|.$$

Example 2.1.2 (Square transformation). Let $Y = X^2$ for X continuous with pdf f_X, so that

$$F_Y(y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

Differentiating,

$$f_Y(y) = \frac{1}{2\sqrt{y}}[f_X(\sqrt{y}) + f_X(-\sqrt{y})].$$

Applying this to standard normal X, we get the chi-square

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}}e^{-y/2}.$$

Here, x^2 is monotonic after partitioning $\mathbb{R}$ into the positive and negative numbers. The book's Theorem 2.1.8 shows that this intuition holds for any function that can be broken up over intervals such that it is monotonic on each interval.

Theorem 2.1.3 (pdfs after transformations which are direct sums of monotinic functions.). *Let X have pdf f_X and $Y = g(X)$. Suppose that $A_0, A_1, \ldots, A_k$ partition $\mathcal{X}$ such that $P(X \in A_0) = 0$ and f_X is continuous on A_i. Further, say that $g(x)$ can be broken up into monotonic functions $g_i(x)$ on each A_i for $i > 0$.*

Definition 2.1.1 (Quantile function). This stands in for inverses of cdfs everywhere, with

$$F_X^{-1}(y) = \inf\{x : F_X(x) \geq y\}.$$

Theorem 2.1.4 (Probability integral transformation). *If F_X is continuous, then $Y = F_X(X)$ is uniform on $(0,1)$.*

Schol: If F_X is not invertible, then use the quantile function.

2.2 Expectation

This section defines the expectation

$$E[g(X)] = \int_{\mathcal{X}} g(x)\, d\mu(x),$$

and notes basic properties: linearity and monotonicity.

2.3 Moment generating functions

Definition 2.3.1 (Moment). Recall the non-central nth moment

$$\mu'_n = EX^n$$

and the nth central moment

$$\mu_n = E(X - \mu)^n,$$

where $\mu = \mu'_1$.

Definition 2.3.2 (Variance). Write $\text{var}(X) = E(X-\mu)^2 = E[X^2] - \mu^2$.

Definition 2.3.3 (Moment generating function). Let X have cdf F_X, and define the mgf

$$M_X(t) = Ee^{tX}$$

as long as it exists in a neighborhood of $t = 0$.

Theorem 2.3.1 (Not a misnomer). *If X has mgf M_X, then*

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

Theorem 2.3.2 (Uniqueness and continuity). *Let F_X, F_Y be cdfs with moments of all orders.*

a. *If X and Y have bounded support, then $F_X = F_Y$ iff $E[X^r] = E[Y^r]$ for all $r = 0, 1, 2....$*

b. *If the mgfs exist and agree on some neighborhood of $t = 0$, then $F_X = F_Y$.*

c. *If $\{X_i\}$ are a sequence of random variables with mgfs M_{X_i} and*

$$\lim_{i\to\infty} M_{X_i}(t) = M_X(t)$$

for all t in some neighborhood of 0 and for some mgf M_X, then there is a unique cdf F_X determined by M_X such that $F_{X_i}(x) \to F_X(x)$ for x where F_X is continuous.

Lemma: If $a_n \to a$, then $(1 + a_n/n)^n \to e^a$.

Theorem 2.3.3 (Shifting and scaling mgf).

$$M_{aX+b}(t) = e^{bt} M_X(at).$$

2.4 Differentiation under the integral sign

Theorem 2.4.1 (Leibniz's rule). *If $f(x,\theta)$, $a(\theta)$, $b(\theta)$ are differentiable wrt θ, then*

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)\,dx = f(b(\theta),\theta)b'(\theta) - f(a(\theta),\theta)a'(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta} f(x,\theta)\,dx.$$

If a and b are constant,

$$\frac{d}{d\theta}\int_a^b f(x,\theta)\,dx = \int_a^b \frac{\partial}{\partial\theta} f(x,\theta)\,dx.$$

Theorem 2.4.2 (Dominated convergence for limits). *Let $h(x,y)$ continuous at y_0 for all x, $|h(x,y)| \le g(x)$ integrable. Then*

$$\lim_{y\to y_0}\int_{-\infty}^{\infty} h(x,y)\,dx = \int_{-\infty}^{\infty} h(x,y_0)\,dx.$$

Theorem 2.4.3 (Dominated convergence for derivatives)**.** *Let $f(x,\theta)$ differentiable in θ, and $|f(x,\theta)| \leq g(x,\theta)$ for $g(x,\theta)$ integrable in x for all θ. Then*

$$\frac{d}{d\theta}\int_{-\infty}^{\theta} f(x,\theta)\,dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f(x,\theta)\,dx.$$

Note: this also holds at just one $\theta = \theta_0$, replacing θ with θ_0 where appropriate.

Theorem 2.4.4 (Differentiating under the sum)**.** *Say that $\sum_{x=0}^{\infty} h(x,\theta)$ converges for all $\theta \in (a,b)$ and*

i. $\frac{\partial}{\partial\theta}h(x,\theta)$ *is continuous in θ for all x,*

ii. $\sum_{x=0}^{\infty} \frac{\partial}{\partial\theta}h(x,\theta)$ *converges uniformly on closed bounded subintervals of (a,b).*

Then,

$$\frac{d}{d\theta}\sum_{x=0}^{\infty} h(x,\theta) = \sum_{x=0}^{\infty} \frac{\partial}{\partial\theta}h(x,\theta).$$

Theorem 2.4.5 (Exchanging integral and sum)**.** *Suppose the series $\sum_{x=0}^{\infty} h(x,\theta)$ converges uniformly for $\theta \in [a,b]$ and for each x, $h(x,\theta)$ is continuous in θ. Then*

$$\int_a^b \sum_{x=0}^{\infty} h(x,\theta)\,d\theta = \sum_{x=0}^{\infty} \int_a^b h(x,\theta)\,d\theta.$$

Chapter 3

Common Families of Distributions

3.1 Discrete distributions

- Discrete uniform on $\{1, \dots, N\}$. Note the mean is the average of the endpoints
$$EX = \frac{1+N}{2}$$
and the variance is
$$\frac{(N+1)(N-1)}{12}.$$
These can both be verified with the identities
$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \quad \text{and} \quad \sum_{i=1}^{n} i^2 = \frac{k(k+1)(k+2)}{6}.$$

- Hypergeometric.
- Bernoulli: $EX = p$, $\operatorname{var} X = p(1-p)$.
- Binomial
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$
To see that this is a pmf, recall the binomial theorem
$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}.$$
Then $EX = np$, $\operatorname{var} X = np(1-p)$.

- Poisson
$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$
To see that this is a pmf, recall the Taylor expansion
$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$
and consider e^λ. It has $EX = \operatorname{var} X = \lambda$. To remember how to simplify complicated binomial expressions, recall that a Poisson process with rate $\lambda = np$ has expectation np.
- Negative binomial.
- Geometric
$$P(X = x) = p(1-p)^{x-1},$$
and you can see from the pmf we have success on the xth trial. To see that it's a pmf, recall
$$\sum_{x=1}^{\infty} q^{x-1} = \frac{1}{1-q}.$$
Then we have mean $EX = \frac{1}{p}$ and variance $\frac{1-p}{p^2}$.

3.2 Continuous distributions

- Uniform$[a, b]$ has pmf $\frac{1}{b-a}$ on its support.
- Gamma has pmf related to the gamma function
$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt,$$
which satisfies $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ and $\Gamma(n) = (n-1)!$. Then clearly
$$f(t) = \frac{t^{\alpha-1} e^{-t}}{\Gamma(\alpha)}$$
is a pmf, but the full gamma(α, β) family has another parameter obtained by the change of variables $x = \beta t$ to add "spread,"" so that
$$f(t \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}.$$
The mean is $EX = \alpha\beta$ and the variance $\operatorname{var} X = \alpha\beta^2$.

There is a relation to Poisson. If X is gamma(α, β), then $P(X \le x) = P(Y \ge \alpha)$ for Y Poisson(x/β).
- The chi squared pmf with p degrees of freedom is
$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2},$$
which is the special case $\alpha = p/2$ and $\beta = 2$ of the gamma.

- Exponential $f(x \mid \beta) = \frac{1}{\beta} e^{-x/\beta}$\$ is another special case of the gamma obtained with $\alpha = 1$. Exponential is the continuous geometric.
- Weibull.
- Normal distribution To perform calculations about the Gaussian with mean μ and variance σ^2, find the standard normal as

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

  and use the linearity properties of your statistic.

To establish that the standard normal density

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

is a pdf, you have to square the integral and go to polar coordinates.

- Beta distribution.
- Cauchy

  $$f(x \mid \theta) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}.$$

  Notice that the cdf is $\frac{1}{2} + \frac{\operatorname{atan}(x-\theta)}{\pi}$. Interestingly it is the pdf of the ratio of two standard normals!
- Lognormal $\log X \sim N(\mu, \sigma^2)$ has pdf

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

  It has expectation $EX = Ee^{\log X} = Ee^{Y} = e^{\mu + \sigma^2/2}$, obtained by using Gaussian mgf.
- Double exponential.

3.3 Exponential families

Definition 3.3.1 (Exponential family)**.** A family of distributions is called exponential if it can be written as

$$f(x \mid \theta) = h(x)c(\theta)exp\Big(\sum_{i=1}^{k} w_i(\theta)t_i(x)\Big).$$

Theorem 3.3.1 (Exponential families let you calculate by differentiation)**.** *If X is drawn from an exponential family, then*

$$E\left[\sum_{i=1}^{k} \frac{\partial}{\partial \theta_j} w_i^{\langle}(\theta)t_i(X)\right] = -\frac{\partial}{\partial \theta_j} \log c(\theta),$$

$$\operatorname{var}\left[\sum_{i=1}^{k} \frac{\partial}{\partial \theta_j} w_i^{\langle}(\theta)t_i(X)\right] = -\frac{\partial^2}{\partial \theta_j^2} \log c(\theta) - E\left[\sum_{i=1}^{k} \frac{\partial^2}{\partial \theta_j^2} w_i^{\langle}(\theta)t_i(X)\right]$$

Curved exponential families.

3.4 Location and scale families

3.5 Probability inequalities

Theorem 3.5.1 (Chebyshev's inequality). *For g nonnegative and $r > 0$,*

$$P(g(X) \geq r) \leq \frac{Eg(X)}{r}.$$

Theorem 3.5.2 (A normal inequality). *Let Z standard normal. Then*

$$P(|Z| \geq t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}.$$

3.6 Identities

There are more than this, but.

Lemma 3.6.1 (Stein's lemma). Let $X \sim N(\mu, \sigma^2)$ and let g have an integrable derivative. Then

$$E[g(X)(X - \mu)] = \sigma^2 Eg'(X).$$

Chapter 4

Multiple Random Variables

4.1 Joint and marginal distributions

Remember,

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} = \frac{\partial^2 P(X \leq x, Y \leq y)}{\partial x \partial y}.$$

4.2 Conditional distributions and independence

Recall the conditional density

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

defined wherever $f_X(x) > 0$ (i.e. if Y absolutely continuous wrt X).

X and Y are independent if

$$f(x,y) = f_X(x) f_Y(y),$$

and actually iff there exist g, h such that $f(x,y) = g(x)h(y)$.

That factorization lets you break up integrals:

$$E[g(X)h(Y)] = Eg(X)Eh(Y).$$

4.3 Bivariate transformations

Transformations of independent variables are independent.

Discrete works the same as before: just find the inverse of the transformation as a set function.

Theorem 4.3.1 (Sums of Poisson)**.** *If $X \sim Poisson(\lambda)$ and $Y \sim Poisson(\theta)$, then $X + Y \sim Poisson(\lambda + \theta)$.*

Definition 4.3.1 (Jacobian in two dimensions)**.** Consider the change of variables

$$u = g_1(x, y) \quad \text{and} \quad v = g_2(x, y),$$

and assume that these are injective, so that we can recover

$$x = h_1(u, v) \quad \text{and} \quad y = h_2(u, v).$$

Then the *Jacobian matrix* is the matrix of first partial derivatives of these inverse mappings

$$\begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix},$$

where

$$\begin{aligned} \frac{\partial x}{\partial u} &= \frac{\partial h_1(u, v)}{\partial u} \\ \frac{\partial x}{\partial v} &= \frac{\partial h_1(u, v)}{\partial v} \\ \frac{\partial y}{\partial u} &= \frac{\partial h_2(u, v)}{\partial u} \\ \frac{\partial y}{\partial v} &= \frac{\partial h_2(u, v)}{\partial v}. \end{aligned}$$

We will be interested in the *Jacobian determinant*

$$J = \det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}.$$

Theorem 4.3.2 (Transformations of pdfs in two dimensions)**.** *Let*

$$U = g_1(X, Y) \quad \textit{and} \quad V = g_2(X, Y),$$

and h_1, h_2 give x and y as functions of u and v, as above. Then

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v)) \, |\det J|$$

on the support of U and V, and 0 elsewhere.

4.4 Hierarchical models and mixture distributions

Some identities like $EX = E[E[X \mid Y]]$ and constructions of hierarchical models...

Theorem 4.4.1 (Conditional variance identity)**.** *If the expectations exist,*

$$\operatorname{var} X = E[\operatorname{var}(X \mid Y)] + \operatorname{var}(E[X \mid Y]).$$

4.5 Covariance and correlation

Definition 4.5.1 (Covariance)**.** This is

$$\operatorname{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Theorem 4.5.1 (Another way to write the covariance)**.**

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

Definition 4.5.2 (Correlation)**.** This is

$$\rho_{XY} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Theorem 4.5.2 (Properties of correlation)**.** ρ_{XY} *satisfies*

- $-1 \le \rho_{XY} \le 1$, *and*
- $|\rho_{XY}| = 1$ *iff there exist* $a \neq 0$, b *such that* $Y = aX + b$ *almost surely. Then* $\operatorname{sgn} a = \rho_{XY}$.

Theorem 4.5.3 (Independence and correlation)**.** *If* $X \perp\!\!\!\perp Y$, *then* $\operatorname{cov}(X, Y) = \rho_{XY} = 0$.

Theorem 4.5.4 (Variance of sums of random variables)**.**

$$\operatorname{var}(aX + bY) = a^2 \operatorname{var} X + b^2 \operatorname{var} Y + 2ab \operatorname{cov}(X, Y).$$

Definition 4.5.3 (Bivariate normal)**.** The bivariate normal can be parametrized using the correlation $\rho = \rho_{XY}$, so that

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}}{2(1-\rho^2)} \right).$$

4.6 Multivariate distributions

Definition 4.6.1 (Multinomial coefficient)**.** The number of ways to divide m objects into n groups with k_i objects in each group is

$$\binom{m}{k_1, \dots, k_n} = \frac{m!}{k_1! \cdots k_n!}.$$

Theorem 4.6.1 (Multinomial theorem)**.** *Let* $\mathcal{A}$ *be the multi-indices that sum to* m. *Then*

$$(x_1 + \cdots + x_n)^m = \sum_{k \in \mathcal{A}} \binom{m}{k_1, \dots, k_n} x_1^{k_1} \dots x_n^{k_n}.$$

Definition 4.6.2 (Multinomial distribution)**.** Let p_i be probabilities summing to 1. Then the multinomial pmf over m trials and n "cells" is

$$f(x_1, \dots, x_n) = \binom{m}{x_1, \dots, x_n} p_1^{x_1} \dots p_n^{x_n} = m! \prod_{i=1}^{n} \frac{p_i^{x_i}}{x_i!}.$$

Definition 4.6.3 (Mutual independence)**.** X_i are mutually independent if $f(x_1, dots, x_n) = \prod_{i=1}^{n} f_i(x_i)$.

Note that pairwise independence does not imply mutual independence.

The Jacobian stuff for more than two variables is also recorded here.

4.7 Inequalities

Theorem 4.7.1 (Hölder's inequality)**.** *Let X, Y be random variables and let p, q satsify $p^{-1} + q^{-1} = 1$. Then*

$$|EXY| \leq E|XY| \leq (E|X|^p)^{\frac{1}{p}} (E|Y|^q)^{\frac{1}{q}}.$$

Theorem 4.7.2 (Cauchy-Schwarz)**.**

$$|EXY| \leq E|XY| \leq (E|X|^2)^{\frac{1}{2}} (E|Y|^2)^{\frac{1}{2}}.$$

This implies

$$(\operatorname{cov}(X, Y))^2 \leq \sigma_X^2 \sigma_Y^2.$$

Theorem 4.7.3 (Minkowski's inequality)**.** *Let $1 \leq p < \infty$. Then*

$$(E|X + Y|^p)^{\frac{1}{p}} \leq (E|X|^p)^{\frac{1}{p}} + (E|Y|^p)^{\frac{1}{p}}.$$

Theorem 4.7.4 (Jensen's inequality)**.** *Let g convex. Then*

$$g(EX) \leq E[g(X)].$$

One can apply Jensen to show that the harmonic mean is bounded by the geometric mean is bounded by the arithmetic mean.

Chapter 5

Properties of a Random Sample

5.1 Basic concepts of random samples

A statistic is defined as a function of the data. For example, the sample mean and variance

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}).$$

Theorem 5.1.1 (Properties of sample mean and variance)**.** *For any x_i,*

- *The sample mean minimizes sum of squared errors.*
- $(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1^n} x_i^2 - n\bar{x}^2$.

For iid random variables $X, X_1, \ldots, X_n$ and any function g such that $Eg(X)$ and $\operatorname{var}(g(X))$ *exist,*

- $E\left[\sum_{i=1}^{n} g(X_i)\right] = nEg(X)$,
- $\operatorname{var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\operatorname{var}(g(X))$,

where the independence was only used for the second bullet. Finally, if $EX = \mu$ and $\operatorname{var} X = \sigma^2 < \infty$*, then*

- $E\bar{X} = \mu$,
- $\operatorname{var}\bar{X} = \frac{\sigma^2}{n}$*, and*
- $ES^2 = \sigma^2$.

Theorem 5.1.2 (pdf and mgf of the sample mean)**.** *If X_i are iid,*

- $f_{\bar{X}}(x) = nf(nx)$, *and*
- $M_{\bar{X}}(t) = (M_X(t/n))^n$.

Theorem 5.1.3 (Sum-convolution)**.** *If $X \perp\!\!\!\perp Y$, then*

$$f_{X+Y}(z) = (f_X * f_Y)(t) = \int_{-\infty}^{\infty} f_X(t) f_Y(z-t)\, dt.$$

There is a theorem given regarding the distribution of the summary statistics of exponential families, which is again an exponential family.

5.2 Sampling from the normal distribution

Theorem 5.2.1 (Gaussian sample mean and variance)**.** *Let $X_i \sim_{iid} N(\mu, \sigma^2)$. Then*

- $\bar{X} \perp\!\!\!\perp S^2$,
- $\bar{X} \sim N(\mu, \sigma^2/n)$, *and*
- $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Theorem 5.2.2 (Facts about the)**.** *If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$. If $X_i \sim \chi^2_{p_i}$, then $X_1 + \cdots + X_n \sim \chi^2_{p_1+\cdots+p_n}$.*

There is a theorem about Gaussians to the effect that decorrelation implies independence.

Definition 5.2.1 (Student's)**.** Note that if $X_i \sim_{\text{iid}} N(\mu, \sigma^2)$, then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

are the quotient of independent $N(0, \sigma^2)$ and square root of χ^2_{n-1} random variables.

If $T \sim t_p$, the Student's t distribution with p degrees of freedom, then

$$F_T(t) = \frac{\Gamma((p+1)/2)}{\Gamma(p/2)} \frac{1}{\sqrt{p\pi}} \frac{1}{(1 + t^2/p)^{(p+1)/2}}.$$

Properties:

- Note that this is Cauchy for $p = 1$.
- $ET = 0$ (if $p > 1$).
- $\operatorname{var} T = \frac{p}{p-2}$ (if $p > 2$).
- It has moments up to order $p - 1$.

Definition 5.2.2 (Snedecor's)**.** Let $X_1, \ldots, X_{p+1}$, $Y_1, \ldots, Y_{q+1}$ be two independent random samples from $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$. Then the quotient

$$F = \frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

is positive and has Snedecor's F distribution with p and q degrees of freedom, with pdf

$$f_F(x) = \frac{\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)}\sqrt{\frac{p}{q}}\frac{x^{(p/2)-1}}{(1+(p/q)x)^{(p+q)/2}}.$$

This can also appear as ratios of other variances.

Proposition 5.2.1 (Properties of and)**.** Let $F \sim F_{p,q}$ and $T \sim t_p$. Then,

- $1/F \sim F_{q,p}$,
- $T^2 \sim F_{1,q}$,
- $(p/q)F/(1+(p/q)X) \sim \text{beta}(p/2, q/2)$.

5.3 Order statistics

Definition 5.3.1 (Order statistics)**.** The order statistics of a random sample $X_1, \ldots, X_n$ are the sorted values of the sample, denoted $X_{(1)}, \ldots, X_{(n)}$.

The rest of the section is more or less devoted to deriving the pdf of the order statistics.

Theorem 5.3.1 (pdf of order statistics)**.** *Let X_i have cdf and pdf F_X,f_X. Then*

$$f_{X_{(j)}} = \frac{n!}{(j-1)!(n-j)!}f_X(x)(F_X(x))^{j-1}(1-F_X(x)^{n-j}).$$

5.4 Convergence concepts

Definition 5.4.1 (Convergence in probability)**.** A sequence of random variables $X_1, X_2, \ldots$ is said to converge in probability to a random variable X if for all $\epsilon > 0$,

$$P(|X_n - X| \geq \epsilon) \to 0.$$

Theorem 5.4.1 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be iid with $EX_i = \mu$, [$\text{var}\, X_i = \sigma^2$ - optional]. Then $\bar{X}_n \to_P \mu$.*

Theorem 5.4.2 (Continuity of)**.** *Let $X_n \to_P X$ and h continuous. Then $h(X_n) \to_P h(X)$.*

Definition 5.4.2 (consistency)**.** An estimator for a parameter is called consistent if it converges in probability to that parameter.

Note that S^2 is consistent in addition to $\bar{X}$ as shown in WLLN.

Definition 5.4.3 (Almost sure convergence)**.** We say that $X_n \to_{\text{a.s.}} X$ if for all $\epsilon > 0$,

$$P(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1.$$

This implies convergence in probability, but not conversely.

Theorem 5.4.3 (Strong law of large numbers). *Let X_i iid with $EX_i = \mu$ [and* $\operatorname{var} X_i = \sigma^2 < \infty$ *- optional]. Then $\bar{X}_n \to_{a.s.} \mu$.*

Definition 5.4.4 (Convergence in distribution). A sequence of random variables $X_1, X_2, \ldots$ converges in distribution to X if

$$F_{X_n}(x) \to F_X(x)$$

at all continuity points of F_X.

Convergence in probability implies weak convergence but not conversely, unless the limit is a constant, in which case the converse holds.

Theorem 5.4.4 (Central limit theorem). *Let $X_1, X_2, \ldots$ be iid with $EX_i = \mu$ and* $\operatorname{var} X_i = \sigma^2$. *Then*

$$\frac{S_n - \mu}{\sqrt{n}\sigma} \Rightarrow N(0,1).$$

Theorem 5.4.5 (Slutsky's theorem). *If $X_n \Rightarrow X$ and $Y_n \to_P a$, then*

- $Y_n X_n \Rightarrow aX$,
- $X_n + Y_n \Rightarrow X + a$.

Definition 5.4.5 (Taylor). Recall the Taylor expansion of g of order r about a,

$$g(x) \approx \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i.$$

Definition 5.4.6 (Approximately normal). We say that $Y_n \sim AN(\mu, \sigma^2)$ if $\sqrt{n}(Y_n - \mu) \Rightarrow N(0, \sigma^2)$.

Theorem 5.4.6 (Delta method). *Let $Y_n \sim AN(\mu, \sigma^2)$. Let g be differentiable at μ and $g'(\mu) \neq 0$. Then $g(Y_n) \sim AN(g(\mu), \sigma^2(g'(\mu))^2)$.*

Theorem 5.4.7 (Second-order delta method). *Let $Y_n \sim AN(\mu, \sigma^2)$. Let g be such that $g'(\mu) = 0$ and $g''(\theta) \neq 0$. Then $n(g(Y_n) - g(\mu)) \Rightarrow \sigma^2 \frac{g''(\mu)}{2} \chi_1^2$.*

Theorem 5.4.8 (Multivariate delta method). *Let $\mathbf{X}_1, \ldots \mathbf{X}_n$ satisfy $E\mathbf{X} = \boldsymbol{\mu}$ and* $\operatorname{cov}(\mathbf{X}, \mathbf{X}) = \Sigma$. *Then for g with continuous first partial derivatives such that*

$$\tau^2 = \sum_i \sum_j \Sigma_{ij} \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_i} \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_j} > 0,$$

it holds that

$$\sqrt{n}(g(\bar{\mathbf{X}}) - g(\boldsymbol{\mu})) \Rightarrow N(0, \tau^2),$$

or in other words, $g(\bar{\mathbf{X}}) \sim AN(g(\boldsymbol{\mu}), \tau^2)$.

5.5 Generating a random sample

This covers inverse CDF method, Box-Muller, rejection sampling, and states Metropolis.