STATS200 class notes

Erich Trieschman

2021 Fall quarter

1 Review: Combinatorics and probability

1.1 Calculus cheat sheet

1.1.1 Logs

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

1.1.2 Derivatives

- $(x^n)' = nx^{n-1}$
- $(e^x)' = e^x$
- $(e^{u(x)})' = u'(x)e^x$
- $(log_e(x))' = (lnx)' = \frac{1}{x}$
- $(f(g(x)))' = f'(g(x))g'(x)$

1.1.3 Integrals

- TODO: Integration by parts

1.1.4 Infinite series and sums

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

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: $(E \cup F) \cap G = (E \cap G) \cup (F \cap G) = E \cap G \cup F \cap G$ and $E \cap F \cup G = (E \cup G) \cap (F \cup G) = E \cup G \cap F \cup G$

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 θ-field (below)
- P : Probability for each event

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

- $0 \in \mathcal{E}$
- $E \in \mathcal{E} \Rightarrow E^C \in \mathcal{E}$
- $E_i \in \mathcal{E}$ for $1, 2, \ldots \Rightarrow \cup_{i=1}^{\infty} E_i \in \mathcal{E}$

Basic probability properties

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

The **law of total probability** relates marginal probabilities to conditional probabilities. For a partition, $E_1, E_2, \ldots$ 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

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

The **continuity of probability measures** state

$$\begin{array}{l} (i)\ E_1 \subset E_2 \subset \ldots \text{ Let } E_\infty = \cup_i E_i, \text{ then } P(E_n) \longrightarrow P(E_\infty) \text{ as } n \longrightarrow \infty \\ (ii)\ E_1 \supset E_2 \supset \ldots \text{ Let } E_\infty = \cap_i E_i, \text{ then } P(E_n) \longrightarrow P(E_\infty) \text{ as } n \longrightarrow \infty \end{array}$$

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 unmeasured events:

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

Where $E_1, E_2, \ldots$ 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 expected value

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 ($Bernouli(p)$): TODO

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

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:** $\frac{1-p}{p}$

2.1.4 Negative binomial

Probability mass function ($NB(r,p)$): TODO

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

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

2.1.5 Poisson distribution

Probability mass function ($Pois(\lambda)$): TODO

$$\frac{\lambda^k e^{-\lambda}}{k!}$$

Expected value: λ **Variance:** λ

2.1.6 Hypergeometric distribution

Probability mass function ($todo$): TODO

Expected value: $todo$ **Variance:** $todo$

2.2 Continuous distribution functions

2.2.1 Uniform distribution

Probability density function (*todo*):

Expected value: *todo* **Variance:** *todo*

2.2.2 Normal distribution

Probability density function (*todo*):

Expected value: *todo* **Variance:** *todo*

2.2.3 Exponential distribution

Probability density function (*todo*):

Expected value: *todo* **Variance:** *todo*

2.2.4 Gamma distribution

Probability density function (*todo*):

Expected value: *todo* **Variance:** *todo*

3 Marginal, joint, and conditional distributions

3.1 Joint distributions

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

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

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

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

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

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

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

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

Likewise, the density of the sum is

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

3.1.2 Expectation of joint distributions

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

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

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

3.2 Marginal distributions

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

$$f_Y(y) = \sum_x yP(Y = y \mid x)$$

3.3 Conditional distributions

Reminder:

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

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

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

4 Expected variables

4.1 Expected value

The expected value (or mean) of a discrete random variable, X, is

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

Which can also be written as

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

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 latter equation structure helps build intuition about the linearity of the expected value function and allows us to derive several other properties of expected values. In the general case:

$$E(g(X)) = \sum_i g(x_i) p_X(x_i), \text{ assuming } g(x_i) = y_i$$

Proof:

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

And from this general equation we can get two key properties of the expected value:

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

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

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

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

5 Variance, covariance, and correlation

The **variance** of X is defined in relation to $E(X) = \mu$ as the expected value of the squared difference between the random variable the mean. The standard deviation, σ is defined as the square root of the variance.

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

Several properties of variance follow from linearity of expectation:

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

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

(ii) $Var(aX + b) = a^2 Var(X)$

$$Var(aX + b) = E((aX + b)^2) - E(aX + b)^2 = E(a^2X^2 + 2abX + b^2) - (aE(X) + b)^2$$
$$Var(aX + b) = a^2E(X^2) + 2abE(X) + b^2 - a^2E(X)^2 - 2abE(X) - b^2 = a^2E(X^2) - a^2E(X)^2 = a^2(E(X^2) - E(X)^2)$$

6 Moment generating functions

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

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

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

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

6.1 Common MGF derivations

7 Convergence and limit theorems

7.1 Convergence in probability

7.2 Convergence in L_p

7.3 Convergence in distribution

7.4 Law of large numbers

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

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

Proof:

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

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

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

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

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

7.5 Central limit theorem

Most useful form of CLT:

$$\sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma} \longrightarrow N(0,1)$$
$$\sqrt{n}(\overline{X}_n - \mu) \longrightarrow N(0,\sigma^2)$$

More formal definition and proof: For $X_1, X_2, \ldots, X_n$ 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

$$\text{For } S_n = \sum_{i=1}^{n} X_i$$
$$\lim_{n\to\infty} P(\frac{S_n}{\sigma\sqrt{n}} \leq x) = \Phi(x)$$

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 $M(s) = M(0) + sM'(0) + \frac{1}{2}sM''(0) + \epsilon_s$

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

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

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

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