STAT 460/560 Class 2: Expectation and inequalities

Ben Bloem-Reddy

Reading: Chapters 3-4, [Was04]

Note: Chapter 3 should be (almost) entirely review; some things in Chapter 4 may be new to you (and it will be useful in this class) so we'll spend more time on that material.

1. Expectation

The expectation of a random variable X with CDF F_X is

$$\mathbb{E}(X) = \int x dF_X(x) = \begin{cases} \sum_x x f_X(x) \,, & \text{if } X \text{ is discrete;} \\ \int x f_X(x) dx \,, & \text{if } X \text{ is continuous.} \end{cases} \tag{2.1}$$

This is well defined as long as $\int |x| dF_X(x) < \infty$. If that is not the case then we say that $\mathbb{E}(X)$ does not exist. Note that Wasserman uses μ to denote $\mathbb{E}(X)$; I will not use that notation because it is very easy to confuse a generic parameter μ with $\mathbb{E}(X)$.

If $Y = r(X)$ then

$$\mathbb{E}(Y) = \mathbb{E}(r(X)) = \int r(x) dF_X(x) \,. \tag{2.2}$$

This is known to statisticians as "The Rule of the Lazy Statistician," and to non-statisticians as "The Rule of the Unconscious Statistician." Take that as you will.

Intuitively, we can think about expectation $\mathbb{E}(r(X))$ as the average $\sum_{i=1}^n r(X_i)/n$ for a large number of IID samples $X_1, \ldots, X_n \sim F_X$. The law of large numbers (next class) justifies this intuition, and lets us use samples from F_X to estimate expectations even when we don't know F_X or an integral of interest in closed form. The technique of using samples to estimate an integral is called Monte Carlo integration.

Exercise 2.1. [Was04], Exercise 3.11.

Also plot the rate of return R_i on an initial investment of $X_0 = 100$:

$$R(X_i) = \frac{X_i - X_0}{X_0} \,. \tag{2.3}$$

2. Moments and moment generating functions

The *k-th moment* of X is $\mathbb{E}(X^k)$, and the *k-th central moment* of X is defined as $\mathbb{E}((X - \mathbb{E}(X))^k)$.

Another useful function is the *moment generating function* (MGF), defined as

$$\psi_X(t) = \mathbb{E}(e^{tX}) \,, \quad t \in \mathbb{R} \,. \tag{2.4}$$

As long as the MGF exists (i.e., is not infinite) in an open interval around $t = 0$ then we can interchange the expectation (integration) with differentiation, so that

$$\left.\frac{d\psi_X(t)}{dt}\right|_{t=0} = \mathbb{E}\left(\frac{d}{dt} e^{tX}\right)_{t=0} = \mathbb{E}(X e^{tX})_{t=0} = \mathbb{E}(X) \,. \tag{2.5}$$

That is, the first derivative of the MGF with respect to t yields the first moment of X. This generalizes to arbitrary (integer) k, so that $\psi_X^{(k)}(0) = \mathbb{E}(X^k)$.

Activity 2.1. Suppose that $X_1, \ldots, X_n$ are independent and each has well-defined MGF ψ_{X_i}, respectively. Let $a_i, b_i \in \mathbb{R}$ for each $i = 1, \ldots, n$, and define $Y = \sum_{i=1}^n a_i X_i + b_i$. Show that

$$\psi_Y(t) = \prod_{i=1}^n e^{b_i t} \psi_{X_i}(a_i t) \, . \tag{2.6}$$

Solution: We have

$$\begin{aligned}
\psi_Y(t) &= \mathbb{E}\left[e^{t\sum_{i=1}^n a_i X_i + b_i}\right] && \text{(definition)} \\
&= \mathbb{E}\left[\prod_{i=1}^n e^{t(a_i X_i + b_i)}\right] && \text{(property of exp)} \\
&= \prod_{i=1}^n \mathbb{E}\left[e^{t(a_i X_i + b_i)}\right] && \text{(by independence)} \\
&= \prod_{i=1}^n e^{b_i t} \psi_{X_i}(a_i t) && \text{(algebra, def. of } \psi_{X_i}\text{)}.
\end{aligned}$$

Activity 2.2. Recall the Gamma distribution from Exercise 1.8, with PDF

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \, , \quad x \in (0, \infty) \, . \tag{2.7}$$

Find the MGF of the Gamma distribution.

Now suppose that $X_i \sim \text{Gamma}(\alpha_i, \beta)$ for $i = 1, \ldots, n$ is a sequence of independent random variables. Use the MGF to find the distribution of $Y = \sum_{i=1}^n X_i$.

Solution:

3. Probability inequalities

Perhaps the most widely used probability inequality is *Markov's inequality*, which says that if X is a non-negative random variable whose expectation exists then for any $t > 0$,

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t} \, . \tag{2.8}$$

The relatively weak conditions for Markov's inequality to apply mean that it gets used in a lot of different situations. However, the generality sometimes comes at a cost: Markov's inequality is often quite loose (the right-hand side is relatively large); if more can be assumed about the distribution of X then tighter bounds are usually possible. We will see some examples below.

Markov's inequality gives rise to a number of other named inequalities. They follow from a generalization of Markov's inequality: Suppose that X is $\mathbb{R}$-valued and that $f \colon \mathbb{R} \to \mathbb{R}_+$ is an increasing function such that $\mathbb{E}(f(X))$ exists. Then

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(f(X))}{f(t)} \, , \quad t \in \mathbb{R} \, . \tag{2.9}$$

Two of particular importance:

- *Chebyshev's inequality*: Letting $f(X) = |X - \mathbb{E}(X)|^2$ yields

$$\mathbb{P}(|X - \mathbb{E}(X)| > t) \leq \frac{\mathrm{Var}(X)}{t^2} , \quad t > 0 . \tag{2.10}$$

- *Chernoff's inequality/bound*, which refers to a family of bounds or a general technique for obtaining probability bounds: Let $f(X) = e^{aX}$ for $a > 0$. Then

$$\mathbb{P}(X > t) \leq e^{-at}\psi_X(a) . \tag{2.11}$$

  Since this holds for all $a > 0$, we can minimize with respect to a:

$$\mathbb{P}(X > t) \leq \inf_{a>0} e^{-at}\psi_X(a) . \tag{2.12}$$

  Starting from here, the typical procedure is to further bound $\psi_X(a)$ to get a bound that is relatively easy to compute and/or manipulate algebraically.

  For a bound on the other tail, we can consider $a < 0$, in which case

$$\mathbb{P}(X < t) = \mathbb{P}(e^{aX} > e^{at}) \leq \inf_{a<0} e^{-at}\psi_X(a) . \tag{2.13}$$

Activity 2.3. Let $X_1, \ldots, X_n$ be IID Bern(p) random variables. Denote the sample mean by $\bar{X}_n$. Show that

$$\mathbb{P}(\bar{X}_n - p > t) \leq \inf_{a>0} e^{-an(t+p)} e^{np(e^a-1)} = e^{nt}\left(1 + \frac{t}{p}\right)^{-n(t+p)} . \tag{2.14}$$

Hint: You may find it helpful to recall that $1 + x \leq e^x$ for all $x \in \mathbb{R}$.

Solution: We have

$$\mathbb{P}(\bar{X}_n - p > t) = \mathbb{P}(n\bar{X}_n - np > nt) = \mathbb{P}\left(\sum_{i=1}^{n} X_i > nt + np\right) \tag{2.15}$$

Using Chernoff's bound and the fact that the X_i's are IID,

$$\begin{aligned}
\mathbb{P}(\bar{X}_n - p > t) &\leq \inf_{a>0} e^{-an(t+p)}\psi_{X_1}(a)^n \\
&= \inf_{a>0} e^{-an(t+p)}(pe^a + (1-p))^n \\
&= \inf_{a>0} e^{-an(t+p)}(1 + p(e^a - 1))^n \\
&\leq \inf_{a>0} e^{-an(t+p)} e^{np(e^a-1)}
\end{aligned}$$

We can minimize the bound by minimizing the exponent (since exp is a strictly increasing function). Taking the derivative with respect to a yields

$$\frac{d}{da}\left(-an(t+p) + np(e^a - 1)\right) = -n(t+p) + npe^a = 0 \Rightarrow a^* = \log\left(1 + \frac{t}{p}\right) .$$

This is greater than zero for $t > 0$. The second derivative with respect to a is $npe^a > 0$, which indicates

that a^* corresponds to the unique minimum. Plugging a^* back into the bound, we have

$$\begin{aligned}\mathbb{P}(\bar{X}_n - p > t) &\leq \exp\left(-n(t+p)\log\left(1+\frac{t}{p}\right) + np(1+t/p-1)\right) \\ &= e^{nt}\left(1+\frac{t}{p}\right)^{-n(t+p)}\end{aligned}$$

A similar analysis of the left tail yields

$$\mathbb{P}(\bar{X}_n - p < -t) \leq \inf_{a<0} e^{-an(p-t)}e^{np(e^a-1)} \tag{2.16}$$

$$= e^{-nt}(1-t/p)^{-n(p-t)} \, . \tag{2.17}$$

This can be combined with the bound from Activity 2.3 to get the two-sided bound

$$\mathbb{P}(|\bar{X}_n - p| > t) \leq \mathbb{P}(\bar{X}_n - p > t) + \mathbb{P}(\bar{X}_n - p < -t) \tag{2.18}$$

$$\leq e^{nt}\left(1+t/p\right)^{-n(t+p)} + e^{-nt}(1-t/p)^{-n(p-t)} \, . \tag{2.19}$$

This is hard to work with algebraically if we want to do anything more, but numerical methods could be used if needed.

If we know or assume more about the distribution then we can get even sharper inequalities. *Hoeffding's inequality*, when applicable, is often about as sharp as we can hope for. Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}(X_i) = 0$ and $\mathbb{P}(a_i \leq X_i \leq b_i) = 1$ for each $i = 1, \ldots, n$. Then for any $a > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \leq \exp\left(-at + \frac{a^2}{8}\sum_{i=1}^n (b_i - a_i)^2\right) \, . \tag{2.20}$$

The proof of this starts with Chernoff's inequality, then uses the bounded support of X_i to get a sharp bound on the MGF. Wasserman has a proof in the Appendix of Chapter 4.

This can be specialized to the mean of an IID sequence of Bern(p) random variables: Let $Y_i = (X_i - p)/n$. Then $\mathbb{E}(Y_i) = 0$, $a_i = -p/n$, and $b_i = (1-p)/n$ for each i. Observe that $(b_i - a_i)^2 = 1/n^2$. Applying Hoeffding's inequality yields, for any $a > 0$,

$$\mathbb{P}(\bar{X}_n - p > t) = \mathbb{P}(\sum_{i=1}^n Y_i > t) \leq e^{-at}e^{a^2/8n} \, . \tag{2.21}$$

Minimizing this with respect to a indicates that we should set $a = 4nt$, in which case

$$\mathbb{P}(\bar{X}_n - p > t) \leq e^{-2nt^2} \, . \tag{2.22}$$

The lower tail yields the same bound, so the two-sided bound is

$$\mathbb{P}(|\bar{X}_n - p| > t) \leq 2e^{-2nt^2} \, . \tag{2.23}$$

Exercise 2.2. For an IID sequence of Bern(p) random variables, obtain a two-sided bound using Chebyshev's inequality.

Plot the three two-sided bounds as a function of n (varying from 10 to 1000) for $t = 0.05$ for each of $p \in \{0.1, 0.25, 0.5, 0.9\}$. What do you notice?

4. Confidence intervals from two-sided bounds

If we're lucky, a two-sided bound can be inverted to obtain a sequence of t_n's so that the probability bound is a constant, say α. For example, using the Hoeffding bound of the Bernoulli sample mean, setting

$$t_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)} \tag{2.24}$$

yields $\mathbb{P}(|\bar{X}_n - p| > t_n) \leq \alpha$, and therefore $(\bar{X}_n - t_n, \bar{X}_n + t_n)$ is a $1 - \alpha$ confidence interval.