STAT 801: Mathematical Statistics

Expectation, moments

Two elementary definitions of expected values:
Defn: If X has density f then

$$E(g(X)) = \int g(x)f(x)\,dx\,.$$

Defn: If X has discrete density f then

$$E(g(X)) = \sum_x g(x)f(x)\,.$$

FACT: If $Y = g(X)$ for a smooth g

$$\begin{aligned} E(Y) &= \int y f_Y(y)\,dy \\ &= \int g(x) f_Y(g(x)) g'(x)\,dy \\ &= E(g(X)) \end{aligned}$$

by the change of variables formula for integration. This is good because otherwise we might have two different values for $E(e^X)$.

In general, there are random variables which are neither absolutely continuous nor discrete. Here's how probabilists define E in general.

Defn: RV X is simple if we can write

$$X(\omega) = \sum_1^n a_i 1(\omega \in A_i)$$

for some constants $a_1, \ldots, a_n$ and events A_i.

Defn: For a simple rv X define

$$E(X) = \sum a_i P(A_i)$$

For positive random variables which are not simple we extend our definition by approximation:
Defn: If $X \ge 0$ then

$$E(X) = \sup\{E(Y) : 0 \le Y \le X, Y \text{ simple}\}$$

Defn: We call X **integrable** if

$$E(|X|) < \infty\,.$$

In this case we define

$$E(X) = E(\max(X,0)) - E(\max(-X,0))\,.$$

Facts: E is a linear, monotone, positive operator:

1. **Linear**: $E(aX + bY) = aE(X) + bE(Y)$ provided X and Y are integrable.
2. **Positive**: $P(X \ge 0) = 1$ implies $E(X) \ge 0$.
3. **Monotone**: $P(X \ge Y) = 1$ and X, Y integrable implies $E(X) \ge E(Y)$.

Major technical theorems:
Monotone Convergence: If $0 \le X_1 \le X_2 \le \cdots$ and $X = \lim X_n$ (which has to exist) then

$$E(X) = \lim_{n\to\infty} E(X_n)\,.$$

Dominated Convergence: If $|X_n| \le Y_n$ and $\exists$ rv X such that $X_n \to X$ (technical details of this convergence later in the course) and a random variable Y such that $Y_n \to Y$ with $E(Y_n) \to E(Y) < \infty$ then

$$E(X_n) \to E(X) .$$

Often used with all Y_n the same rv Y.

Fatou's Lemma: If $X_n \ge 0$ then

$$E(\limsup X_n) \le \limsup E(X_n) .$$

Theorem: With this definition of E if X has density $f(x)$ (even in R^p say) and $Y = g(X)$ then

$$E(Y) = \int g(x) f(x) dx .$$

(Could be a multiple integral.) If X has pmf f then

$$E(Y) = \sum_x g(x) f(x) .$$

First conclusion works, e.g., even if X has a density but Y doesn't.

Defn: The r^{th} moment (about the origin) of a real rv X is $\mu_r' = E(X^r)$ (provided it exists). We generally use μ for $E(X)$.

Defn: The r^{th} central moment is

$$\mu_r = E[(X-\mu)^r] .$$

We call $\sigma^2 = \mu_2$ the variance.

Defn: For an R^p valued random vector X

$$\mu_X = E(X)$$

is vector whose i^{th} entry is $E(X_i)$ (provided all entries exist).

Defn: The $(p \times p)$ variance covariance matrix of X is

$$Var(X) = E\left[(X-\mu)(X-\mu)^t\right]$$

which exists provided each component X_i has a finite second moment.

Moments and probabilities of rare events are closely connected as will be seen in a number of important probability theorems.

Example: Markov's inequality

$$\begin{aligned} P(|X-\mu| \ge t) &= E[1(|X-\mu| \ge t)] \\ &\le E\left[\frac{|X-\mu|^r}{t^r} 1(|X-\mu| \ge t)\right] \\ &\le \frac{E[|X-\mu|^r]}{t^r} \end{aligned}$$

Intuition: if moments are small then large deviations from average are unlikely.

Special Case: Chebyshev's inequality

$$P(|X-\mu| \ge t) \le \frac{\text{Var}(X)}{t^2} .$$

Example moments: If Z is standard normal then

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} z e^{-z^2/2} dz / \sqrt{2\pi} \\ &= \left. \frac{-e^{-z^2/2}}{\sqrt{2\pi}} \right|_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

and (integrating by parts)

$$\begin{aligned}E(Z^r) &= \int_{-\infty}^{\infty} z^r e^{-z^2/2} dz/\sqrt{2\pi} \\ &= \left.\frac{-z^{r-1}e^{-z^2/2}}{\sqrt{2\pi}}\right|_{-\infty}^{\infty} \\ &\quad + (r-1)\int_{-\infty}^{\infty} z^{r-2}e^{-z^2/2}dz/\sqrt{2\pi}\end{aligned}$$

so that

$$\mu_r = (r-1)\mu_{r-2}$$

for $r \ge 2$. Remembering that $\mu_1 = 0$ and

$$\mu_0 = \int_{-\infty}^{\infty} z^0 e^{-z^2/2} dz/\sqrt{2\pi} = 1$$

we find that

$$\mu_r = \begin{cases} 0 & r \text{ odd} \\ (r-1)(r-3)\cdots 1 & r \text{ even}. \end{cases}$$

If now $X \sim N(\mu, \sigma^2)$, that is, $X \sim \sigma Z + \mu$, then $E(X) = \sigma E(Z) + \mu = \mu$ and

$$\mu_r(X) = E[(X-\mu)^r] = \sigma^r E(Z^r).$$

In particular, we see that our choice of notation $N(\mu, \sigma^2)$ for the distribution of $\sigma Z + \mu$ is justified; σ is indeed the variance.

Similarly for $X =\sim MVN(\mu, \Sigma)$ we have $X = AZ + \mu$ with $Z \sim MVN(0, I)$ and

$$E(X) = \mu$$

and

$$\begin{aligned}\text{Var}(X) &= E\left\{(X-\mu)(X-\mu)^t\right\} \\ &= E\left\{AZ(AZ)^t\right\} \\ &= AE(ZZ^t)A^t \\ &= AIA^t = \Sigma.\end{aligned}$$

Note use of easy calculation: $E(Z) = 0$ and

$$\text{Var}(Z) = E(ZZ^t) = I.$$

Moments and independence

Theorem: If $X_1, \ldots, X_p$ are independent and each X_i is integrable then $X = X_1 \cdots X_p$ is integrable and

$$E(X_1 \cdots X_p) = E(X_1) \cdots E(X_p).$$

Proof: Suppose each X_i is simple:

$$X_i = \sum_j x_{ij} 1(X_i = x_{ij})$$

where the x_{ij} are the possible values of X_i. Then

$$
\begin{aligned}
E(X_1 \cdots X_p) & \\
&= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} \\
&\quad \times E(1(X_1 = x_{1j_1}) \cdots 1(X_p = x_{pj_p})) \\
&= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} \\
&\quad \times P(X_1 = x_{1j_1} \cdots X_p = x_{pj_p}) \\
&= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} \\
&\quad \times P(X_1 = x_{1j_1}) \cdots P(X_p = x_{pj_p}) \\
&= \sum_{j_1} x_{1j_1} P(X_1 = x_{1j_1}) \times \cdots \\
&\quad \times \sum_{j_p} x_{pj_p} P(X_p = x_{pj_p}) \\
&= \prod E(X_i) \, .
\end{aligned}
$$

General $X_i > 0$:

Let X_{in} be X_i rounded down to nearest multiple of 2^{-n} (to maximum of n).

That is: if

$$\frac{k}{2^n} \le X_i < \frac{k+1}{2^n}$$

then $X_{in} = k/2^n$ for $k = 0, \ldots, n2^n$. For $X_i > n$ put$X_{in} = n$.

Apply case just done:

$$E(\prod X_{in}) = \prod E(X_{in}) \, .$$

Monotone convergence applies to both sides.

For general case write each X_i as difference of positive and negative parts:

$$X_i = \max(X_i, 0) - \max(-X_i, 0) \, .$$

Apply positive case.