18.440 Probability and Random Variables

Spring 2009

# 18.440 Exam 2 Review Session

The exam will be closed book with no notes (such as a formula sheet) allowed. You can use calculators, but any other electronic devices such as cell phones need to be turned off and stowed away during the exam. If such a device is available, as on your desk or in your hand, during the exam there will be a 20% penalty (the same as for continuing to write after "stop").

These notes will indicate the material to be covered, which is the material of problem sets 4-6, not including Sessions 23 and 24 which relate to PS7 material.

Beside the Ross textbook, there have been handouts on Stirling's formula and on gamma and beta probabilities. Relevant parts of those are incorporated in these notes.

## 1. Random Variables

A (real-valued) random variable is a function $X$ defined on the sample space $X$ such that for each real number $t$, the probability $F(t) = P(X \leq t)$ is defined. Here $F$ is called the (cumulative) *distribution function* of the random variable. It has the following characteristic properties:

1. $F$ is nondecreasing: if $u \leq x$ then $F(u) \leq F(x)$;
2. Limits at $\pm\infty$: $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = 1$.
3. $F$ is right-continuous: for all $x$, $\lim_{v\downarrow x} F(v) = F(x)$.

## 2. Discrete random variables

A random variable $X$ is called *discrete* if there is a sequence (finite or infinite) $\{x_j\}$ of distinct values such that $\sum_j P(X = x_j) = 1$. For any discrete random variable $X$, its *probability mass function* is the function $f$ such that $f(x) = f_X(x) = P(X = x)$ for all $x$. This will be 0 except when $x$ is one of the possible values $x_j$ of $X$.

The *expectation* of a discrete random variable $X$ is defined as

$$EX \;=\; \sum_x x f_X(x) \;=\; \sum_j x_j P(X = x_j)$$

provided the sum is absolutely convergent, that is $\sum_j |x_j| P(X = x_j) < +\infty$.

For a function $g$ of the discrete random variable $X$ we have the following useful way of finding $Eg(X)$. (It's Proposition 5.1 in Section 4.5 of Ross.)

**Theorem 1.** *(Law of the unconscious statistician — discrete case). Let $X$ be a discrete random variable with possible values $x_j$ and probability mass function $f$. Then for any function $g$,*

$$Eg(X) \;=\; \sum_x g(x) f(x) \;=\; \sum_j g(x_j) f(x_j)$$

*if the series is absolutely convergent, $\sum_j |g(x_j)| f(x_j) < \infty$.*

*Proof.* $Y = g(X)$ is a discrete random variable with some possible values $y_k$ and $P(Y = y_k) = \sum_{j:\ g(x_j)=y_k} f(x_j)$ for each $k$. Thus

$$EY \;=\; \sum_k y_k \sum_{j:\ g(x_j)=y_k} f(x_j) \;=\; \sum_j f(x_j) g(x_j)$$

where after interchanging sums, there is just one $k$ for each $j$. The interchange is justified by first considering $|g(x_j)|$ where all terms will be nonnegative. □

The *variance* of any random variable $X$ (discrete or not) is defined as

$$\mathrm{Var}(X) \;=\; E((X-EX)^2) \;=\; E(X^2)-(EX)^2.$$

Note that $E(X^2)=(EX)^2$, in other words $\mathrm{Var}(X)=0$, if and only if the nonnegative random variable $(X-EX)^2$ has expectation 0, which implies that it equals 0, i.e. $X=EX$, with probability 1, in other words $X$ is a constant. So for non-constant random variables we will have $\mathrm{Var}(X)>0$ and $(EX)^2<E(X^2)$.

For any random variable $X$ and constants $a,b$, $\mathrm{Var}(a+bX)=b^2\mathrm{Var}(X)$.

## 3. Families of discrete random variables

3.1. **Binomial and Bernoulli random variables.** A random variable $X$ is said to have a binomial$(n,p)$ distribution, where $0\le p\le 1$ and $n$ is a positive integer, if with $q\equiv 1-p$,

$$P(X=k) \;=\; b(k,n,p) \;=\; \binom{n}{k}p^kq^{n-k}$$

for $k=0,1,...,n$, and 0 otherwise. Here $X$ is the number of successes in $n$ independent trials with probability $p$ of success on each. In the special case $n=1$ the variable is called a *Bernoulli*$(p)$ random variable. A binomial$(n,p)$ random variable $X$ can be viewed as $X=\sum_{j=1}^n X_j$ where $X_j=1$ if the $j$th trial is a success and 0 otherwise. Thus each $X_j$ is a Bernoulli$(p)$ random variable. We have $EX=np$ and $\mathrm{Var}(X)=npq$.

3.2. **Poisson random variables.** For $0\le\lambda<\infty$, a random variable $Y$ is said to have a Poisson$(\lambda)$ distribution if for each $k=0,1,...,$

$$P(Y=k) \;=\; p(k,\lambda) \;=\; e^{-\lambda}\lambda^k/k!.$$

Such a $Y$ has $EY=\lambda$ and $\mathrm{Var}(Y)=\lambda$. The *Poisson limit theorem* says that if $n\to\infty$ and $p=p_n\to 0$ in such a way that $np_n\to\lambda$ then the binomial probability $b(k,n,p_n)\to p(k,\lambda)$, as was proved in a lecture. This gives a Poisson approximation to some binomial probabilities. According to the proposed rules given for this course, if $n\ge 20$ and $np^2\le 0.1$ then $b(k,n,p)$ can be approximated by $p(k,np)$. If $p$ is close to 1 one may be able to use the "reverse Poisson" approximation. Namely, if $n\ge 20$ and $nq^2\le 0.1$ then $b(k,n,p)=b(n-k,n,q)$ can be approximated by $p(n-k,nq)$. For example, $b(99,100,0.98)=100(0.98)^{99}(0.02)\doteq 0.27065$. This equals $b(1,100,0.02)$ which has the Poisson approximation $p(1,2)=e^{-2}\cdot 2/1!\doteq 0.27067$. Partly by good luck, this approximation works very well. But to take $p(k,np)=p(99,98)=e^{-98}(98)^{99}/99!\doteq 0.0399$ gives a terrible approximation, from failing to do the needed reversal.

3.3. **Geometric random variables.** In a sequence of independent trials with probability $p$ of success, let $X$ be the number of trials needed to give the first success. Then $P(X=k)=q^{k-1}p$ for $k=1,2,3,....$ Such a random variable has $EX=1/p$ and $\mathrm{Var}(X)=q/p^2$. It's natural that if $p$ is small, it takes longer on average to get a first success, so $EX$ is inversely proportional to $p$. If $p=1$ then $X=1$, a constant, with probability 1, and $q=0$, so the variance of $X$ is 0 as it should be.

3.4. **Hypergeometric random variables.** Suppose given a finite set of $N$ objects, of which $m$ have a property A, for example, "defective." Suppose a random sample of $k$ distinct objects is taken (without replacement) from the $N$, with probability $1/\binom{N}{k}$ for each possible sample. Let $X$ be the number of objects having A in the sample. For each $j = 0, 1, ..., \min(k, m)$, $P(X = j) = \binom{m}{j}\binom{N-m}{k-j}/\binom{N}{k}$. What we did so far with hypergeometric probabilities didn't go beyond that one formula, and of course recognizing when hypergeometric probabilities apply (sampling without replacement from a finite population). We have $EX = km/N = kp$ where $p = m/N$ is the probability of having A. To see this let $X_j = 1$ if the $j$th element of the sample has A and 0 otherwise. Then $X = X_1 + \cdots + X_k$ and $EX_j = m/N$ for each $j$, which implies the given form for $EX$. Here $E(X_1 + \cdots + X_k) = EX_1 + \cdots + EX_k$ which holds for any random variables having finite expectations, including dependent ones such as these. For the variance, however, the dependence matters. As it hasn't appeared in the course so far, it won't be given here.

## 4. Stirling's formula

Two sequences of numbers, $a_n$ and $b_n$, are said to be *asymptotic*, written $a_n \sim b_n$, if $\lim_{n\to\infty} a_n/b_n = 1$. This does not imply that $\lim_{n\ \to\infty}(a_n - b_n) = 0$: for example, $n^2 + n \sim n^2$ but $(n^2 + n) - n^2$ tends to $\infty$ with $n$. But $a_n/b_n \to 1$ is equivalent to $\log(a_n) - \log(b_n) = \log(a_n/b_n) \to 0$.

In some other places $f \sim g$ might be meant in the sense of a rather vague approximation, but the definition above gives a precise meaning. Note that to write $a_n \to b_n$ in place of $a_n \sim b_n$ is nonsense: it's a syntax error to write that as $n \to \infty$, $a_n$ converges to something that itself depends on $n$. And also, if it were interpreted to imply that $a_n - b_n \to 0$ that is wrong in a lot of cases.

**Theorem 2.** *Stirling's formula: as $n \to +\infty$,*

$$n! \sim \frac{n^n}{e^n}\sqrt{2\pi n} \;=\; n^{(n+1/2)}\, e^{-n}\sqrt{2\pi}.$$

*Thus,*

$$\log(n!) - \left[\left(n + \frac{1}{2}\right)\log\, n - n + \frac{1}{2}\log(2\pi)\right] \to 0 \;\; \textit{as } n \to \infty.$$

A proof was given in the handout, but it's rather long and certainly not covered in the course.

Examples were given near the end of the handout showing that as $n$ gets large the difference between $n!$ and its Stirling approximation, $D_n = n! - \frac{n^n}{e^n}\sqrt{2\pi n}$, becomes very large very fast, just not quite as fast as $n!$ itself. For example $60!$ is $8.321 \cdot 10^{81}$ to the given number of places and the difference $D_{60} \doteq 1.155 \cdot 10^{79}$, small*er* by a factor of about 720, but not small at all in an absolute sense.

In a binomial coefficient $\binom{n}{k}$, if $k$ and $n - k$ both become large, so we can use Stirling approximations to all three of $n!$, $k!$ and $(n - k)!$, the powers of $e$ will cancel out.

If on the other hand $k$ remains fixed while $n \to \infty$, so $n - k \to \infty$ also, then for $\binom{n}{k}$ we have the asymptotic form

$$\binom{n}{k} \;=\; \frac{n(n-1)\cdots(n-k+1)}{k!} \;\sim\; \frac{n^k}{k!},$$

as $n - j \sim n$ for each $j = 0, 1, ..., k-1$ and there is a fixed number $k$ of factors. The above form was used in proving the Poisson limit theorem. In this case it's not useful to apply Stirling's formula.

## 5. Continuous random variables

A function $f$ defined for all real $x$ is called a *probability density function* if $f(x) \geq 0$ for all $x$ and $\int_{-\infty}^{\infty} f(x)dx = 1$. A random variable $X$ is said to be continuous with density $f$ if for all $x$, the distribution function

$$F(x) \;=\; P(X \leq x) \;=\; \int_{-\infty}^{x} f(u)du.$$

This implies by the fundamental theorem of calculus that given $F$, we can find $f$ as $f(x) = F'(x)$ at any $x$ where $f$ is continuous. (The densities in the families to be considered are continuous except possibly at one or two points.)

Suppose $X$ is a continuous random variable with density $f$ and distribution function $F$, and let $c > 0$. To find the distribution and density of $cX$ we have for any $x$ that

$$F_{cX}(x) \;=\; P(cX \leq x) \;=\; P(X \leq x/c) \;=\; F(x/c)$$

and so by the chain rule, the density of $cX$ is given by

$$f_{cX}(x) \;=\; \frac{1}{c} f\left(\frac{x}{c}\right) \tag{1}$$

at any $x$ such that $f$ is continuous at $x/c$.

For a continuous random variable, the expectation is defined by

$$EX \;=\; \int_{-\infty}^{\infty} x f(x)dx$$

provided $\int_{-\infty}^{\infty} |x| f(x)dx < +\infty$. Similarly as in the discrete case we have:

**Theorem 3.** *(Law of the unconscious statistician — continuous case). Let $X$ be a continuous random variable with probability density function $f$. Then for any random variable $Y$ which is a function of $X$, $Y = g(X)$, we have*

$$EY = Eg(X) \;=\; \int_{-\infty}^{\infty} g(x) f(x)dx$$

*if the integral is absolutely convergent, $\int_{-\infty}^{\infty} |g(x)| f(x)dx < \infty$.*

In the course a direct proof was given for the special but important case $g(x) = x^2$ needed in defining the variance. If $g$ is a continuous function one might argue by way of approximating $X$ and $g(X)$ by approximations to a finite number of decimal places, but no proof is considered to have been covered in the course.

5.1. **Hazard or failure rate.** For a continuous random variable $X$ with $X \geq 0$ having density $f$ and distribution function $F$, the *hazard rate* or *failure rate* is defined as

$$h(t) \;=\; f(t)/(1-F(t))$$

for all $t \geq 0$ such that $F(t) < 1$. Clearly $h(t) = -\frac{d}{dt}\log(1-F(t))$. Thus $\log(1-F(t)) = -\int_0^t h(u)du$ and

$$1 - F(t) = \exp\left(-\int_0^t h(u)du\right). \tag{2}$$

## 6. Families of continuous distributions

6.1. **Uniform distributions.** If $-\infty < a < b < +\infty$ then the $U[a,b]$ distribution is defined to have density $f(x) = 1/(b-a)$ for $a \leq x \leq b$ and 0 elsewhere. If $X$ has this density then $EX = (a+b)/2$ (which seems rather obvious) and $\mathrm{Var}(X) = (b-a)^2/12$. It's rather intuitive that the variance should be proportional to $(b-a)^2$, but the constant $1/12$ is seen from a calculation (which is easy).

6.2. **Normal distributions.** If $-\infty < \mu < \infty$ and $0 < \sigma < \infty$ then the $N(\mu, \sigma^2)$ density is given by

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

The standard normal or $N(0,1)$ density is then $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$. It was shown that this is a probability density by setting $I = \int_{-\infty}^{\infty} e^{-x^2/2}dx$, then writing

$$I^2 \;=\; \int_{-\infty}^{\infty} e^{-x^2/2}dx \int_{-\infty}^{\infty} e^{-y^2/2}dy,$$

and evaluating the double integral in polar coordinates, which gives $2\pi$. It then follows by simple changes of variables that each $N(\mu, \sigma^2)$ density is indeed a probability density.

A variable $X$ has a $N(\mu, \sigma^2)$ distribution (density) if and only if $Z = (X-\mu)/\sigma$ has a $N(0,1)$ distribution. Via that, probabilities for $X$ can be found by way of probabilities for $Z$. The standard normal distribution function

$$\Phi(x) \;=\; \int_{-\infty}^{x} \phi(u)du$$

cannot be evaluated in closed form but is tabulated in many books including Ross. A copy of the table will be attached to each exam. For $z > 0$ and $Z$ with $N(0,1)$ distribution, $\Phi(-z) \equiv 1 - \Phi(z)$ because $\phi(-z) \equiv \phi(z)$. So the table need only be given for $z \geq 0$. The table in Ross goes up to $z = 3.49$ where $\Phi(3.49) \doteq 0.9998$, close to 1.

6.3. **Exponential distributions.** The function $f(x) = e^{-x}$ for $x \geq 0$ and 0 for $x < 0$ is clearly a probability density. It's called the standard exponential density. A random variable $X$ with this density has $EX = 1$ via an integration by parts. By another integration by parts and Theorem 3 we get $E(X^2) = 2$ and so $\mathrm{Var}(X) = 1$.

If $0 < \lambda < \infty$ and $Y = X/\lambda$ then by (1), $Y$ has density $\lambda e^{-\lambda x}$ for $x \geq 0$ and 0 elsewhere. This is called an exponential density with parameter $\lambda$. It has $EY = 1/\lambda$ and $\mathrm{Var}(Y) = 1/\lambda^2$. Often exponential densities are specified by giving their expectations $EY$ and then $\lambda = 1/EY$.

The distribution function of $Y$ is found by direct integration to be $P(Y \le y) = 1 - e^{-\lambda y}$ for any $y \ge 0$ and 0 for $y < 0$. Thus $P(Y > y) = e^{-\lambda y}$. Exponential distributions have the *memoryless property*, for any $y > 0$ and $h > 0$,

$$P(Y > y + h | Y > y) \;=\; e^{-\lambda(y+h)}/e^{-\lambda y} \;=\; e^{-\lambda h} \;=\; P(Y > h).$$

For an exponential distribution with parameter $\lambda$ the failure rate $h(t)$ equals the constant $\lambda$ for all $t > 0$. This is another form of the memoryless property. It follows from (2) that only exponential distributions have constant failure rate.

6.4. **Gamma distributions.** The *Gamma function* is defined for any $a > 0$ by $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$. The integral is finite because $\int_0^1 x^{a-1} dx = 1/a < \infty$ and $e^{-x}$ becomes small for large $x$ much faster than $x^{a-1} \to +\infty$.

We have $\Gamma(1) = 1$ (integral of the standard exponential density). The gamma function satisfies the recurrence relation $\Gamma(a+1) = a\Gamma(a)$ for all $a > 0$ via an integration by parts. It follows that $\Gamma(2) = 1$ (expectation of a standard exponential variable), $\Gamma(3) = 2$, and in general by induction, for any integer $n \ge 0$, $n! = \Gamma(n+1)$.

For any $a > 0$ we get a probability density $\gamma_a$ by setting $\gamma_a(x) = x^{a-1} e^{-x}/\Gamma(a)$ for $x > 0$ and 0 for $x \le 0$. If $X$ has this density then by the definition of the gamma function $EX = \Gamma(a+1)/\Gamma(a) = a$. Likewise $E(X^2) = \Gamma(a+2)/\Gamma(a) = (a+1)a$, using Theorem 3 and then the recurrence formula twice, and so $\mathrm{Var}(X) = a$.

For any $\lambda$ with $0 < \lambda < \infty$, if $X$ has the $\gamma_a$ density, then by (1), $Y = X/\lambda$ has the density

$$\gamma_{a,\lambda}(x) \;=\; \lambda(\lambda x)^{a-1} e^{-\lambda x}/\Gamma(a) \;=\; \lambda^a x^{a-1} e^{-\lambda x}/\Gamma(a)$$

for $x > 0$ and 0 for $x \le 0$. Clearly, $EY = a/\lambda$ and $\mathrm{Var}(Y) = a/\lambda^2$.

It was proved in class and in a handout (the proof itself is not covered on the exam) that if $X$ and $Y$ are independent, $X$ has $\gamma_{a,\lambda}$ density and $Y$ has $\gamma_{b,\lambda}$ for the same $\lambda$, then $X + Y$ has a $\gamma_{a+b,\lambda}$ density. As a byproduct of the proof, we found for the beta function defined for $a > 0$ and $b > 0$ by

$$B(a,b) \;=\; \int_0^1 x^{a-1}(1-x)^{b-1} dx$$

that $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. If $a$ and $b$ are integers, thus $\ge 1$, then

$$B(a,b) \;=\; \frac{(a-1)!(b-1)!}{(a+b-1)!} \;=\; \frac{1}{(a+b-1)\binom{a+b-2}{a-1}},$$

the reciprocal of an integer.