14.30 Introduction to Statistical Methods in Economics

Spring 2009

14.30 Introduction to Statistical Methods in Economics
Lecture Notes 16

Konrad Menzel

April 9, 2009

1 General Exam Policies

- Exam 2 will be in class next Tuesday, April 14, starting at 9:00 *sharp*
- relevant material: in first place topics we covered since last exam, but of course should feel comfortable with densities, probabilities and other concepts from the first third of the semester
- more text problems than on problem sets, but less tedious calculations
- will hand out normal probability tables with exam, so don't have to bring your own
- essentially same format as in first exam
- bring calculators
- closed books, closed notes
- have about 85 minutes to do exam
- I'll give partial credit, so try to get started with all problems

2 Review

2.1 Functions of Random Variables

General set-up:

- know p.d.f. of X, $f_X(x)$ (discrete or continuous)
- Y is a known function of X, $Y = u(X)$
- interested in finding p.d.f. $f_Y(y)$

The way how we obtain the p.d.f. $f_Y(y)$ depends on whether X is continuous or discrete, and whether the function $u(\cdot)$ is one-to-one. Three methods

1. if X discrete:

$$f_Y(y) = \sum_{\{x:u(x)=y\}} f_X(x)$$

2. 2-step method if X continuous: Step 1: obtain c.d.f. $F_Y(y)$

$$F_Y(y) = P(u(X) \leq y) = \int_{\{x:u(x)\leq y\}} f_X(x)dx$$

Step 2: differentiate c.d.f. in order to obtain p.d.f.:

$$f_Y(y) = \frac{d}{dy}F_Y(y)$$

3. change of variables formula if (a) X continuous, and (b) $u(\cdot)$ is one-to-one:

$$f_Y(y) = f_X(s(y)) \left|\frac{d}{dy}s(y)\right|$$

A few important examples which we discussed were:

- Convolution Formula: if X and Y are independent, then $Z = X + Y$ has p.d.f.

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z-w)f_X(w)dw$$

Note: if densities of X and/or Y are zero somewhere, be careful with integration limits!

- Integral Transformation: if X continuous, then the random variable $Y = F_X(X)$, where $F_X(\cdot)$ is the c.d.f. of X is uniformly distributed.
- Order Statistics: if $X_1, \ldots, X_n$ i.i.d., then kth lowest value Y_k has p.d.f.

$$f_{Y_k}(y) = k \left(\begin{array}{c} n \\ k \end{array} \right) F_X(y)^{k-1} \left(1 - F_X(y)\right)^{n-k} f_X(y)$$

2.2 Expectations

2.2.1 Expectation

Definition of expectation of X

- if X discrete,

$$\mathbb{E}[X] = \sum_x x f_X(x)$$

- if X continuous,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx$$

Important properties of expectations

1. for constant a,

$$\mathbb{E}[a] = a$$

2. for linear function of X, $Y = aX + b$,

$$\mathbb{E}[Y] = a\mathbb{E}[X] + b$$

3. for 2 or more random variables,

$$\mathbb{E}[a_1X_1 + \ldots + a_nX_n + b] = a_1\mathbb{E}[X_1] + \ldots + a_n\mathbb{E}[X_n] + b$$

4. if X and Y are *independent*, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

- expectation is measure of *location* of distribution of X.

- expectation of function $Y = u(X)$ (discrete case: replace integral with sum)

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} u(x) f_X(x) dx$$

- Jensen's Inequality: if $u(\cdot)$ is *convex*, then

$$\mathbb{E}[u(X)] \geq u(\mathbb{E}[X])$$

2.2.2 Variance

Defined as

$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

Measure of *dispersion* of X.

Important properties of variances

1. for a constant a,

$$\text{Var}(a) = 0$$

2. can write variance as

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

3. for a linear function of *independent* random variables $X_1, \ldots, X_n$,

$$\text{Var}(a_1X_1 + \ldots + a_nX_n + b) = a_1^2\text{Var}(X_1) + \ldots + a_n^2\text{Var}(X_n) + b$$

4. more generally for *any* random variables X_1, X_2,

$$\text{Var}(a_1X_1 + a_2X_2) = a_1^2\text{Var}(X_1) + 2a_1a_2\text{Cov}(X_1, X_2) + a_2^2\text{Var}(X_2)$$

2.2.3 Covariance and Correlation

Covariance defined as

$$\text{Cov}(X, Y) = \mathbb{E}[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])]$$

Properties of covariances

$$\begin{aligned}
\text{Cov}(X, X) &= \text{Var}(X) \\
\text{Cov}(X, Y) &= \text{Cov}(Y, X) \\
\text{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
\text{Cov}(aX + b, cY + d) &= ac\text{Cov}(X, Y)
\end{aligned}$$

If X and Y are independent, $\text{Cov}(X,Y) = 0$.

Correlation coefficient defined as

$$\varrho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

$\varrho(X < Y) \in [-1,1]$, and $|\varrho(X,Y)| = 1$ if and only if Y is a deterministic *linear* function of X.

2.2.4 Conditional Expectation

Conditional expectation *random variable* defined by

$$\mathbb{E}[Y|X] = \int_{-\infty}^{\infty} y f_{Y|X}(y|X)dy$$

Two important results on conditional expectations:

- Law of Iterated Expectations
$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$$
- Conditional Variance
$$\text{Var}(Y) = \text{Var}(\mathbb{E}[Y|X]) + \mathbb{E}[\text{Var}(Y|X)]$$

2.3 Special Distributions

2.3.1 Summary

Looked following distributions

- Uniform: $X \sim U[a,b]$ if p.d.f. of X is
$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$
- Binomial: $X \sim B(n,p)$ if p.d.f. of X is
$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x \in \{0,1,\ldots,n\} \\ 0 & \text{otherwise} \end{cases}$$
- Exponential: $X \sim E(\lambda)$ if X has p.d.f.
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$
- Normal: $X \sim N(\mu, \sigma^2)$ if p.d.f. is
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
- Poisson: $X \sim P(\lambda)$ if p.d.f. is
$$f_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{if } x \in \{0,1,2,\ldots\} \\ 0 & \text{otherwise} \end{cases}$$

Should know or be able to calculate mean and variance for each distribution. Also showed relationships between Binomial and Poisson, and Binomial and Normal.

2.3.2 Normal Distribution

Should know how to standardize random variables:

$$Z = \frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}(X)}}$$

I will give you a copy of the tabulated c.d.f. of the standard normal, so should know how to read the tables.

Important results on normal distribution:

1. normal p.d.f. is symmetric about the mean
2. linear functions of normal random variables are again normally distributed: if $X \sim N(\mu, \sigma^2)$, then $Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$.
3. sums of independent normal random variables are normally distributed
4. Central Limit Theorem: standardized sample mean for i.i.d. sample $X_1, \ldots, X_n$ approximately follows a standard normal distribution for large n.

2.4 Asymptotic Theory

2.4.1 General Idea

- always assume i.i.d. sample $X_1, \ldots, X_n$
- only deal with sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- exact value/distribution often hard or even impossible to derive given our knowledge about distribution of X_i
- thought experiment "$n \to \infty$" supposed to give *approximations* for large n

2.4.2 Law of Large Numbers

- Chebyshev's Inequality: for any $\varepsilon > 0$,

$$P(|X - \mathbb{E}[X]| > \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}$$

- Law of Large Numbers: if $X_i, \ldots, X_n$ i.i.d., then for all $\varepsilon > 0$

$$\lim_{n\to\infty} P(|\bar{X}_n - \mathbb{E}[X]| > \varepsilon) = 0$$

- independence assumption important (e.g. correlated event in "wisdom of crowds" example.
- need $\text{Var}(X_i) < \infty$, so LLN doesn't work with very fat tailed distributions.

2.4.3 Central Limit Theorem

- look at distribution of *standardized* sample mean
- Central Limit Theorem: for an i.i.d. sample with $\text{Var}(X_i) < \infty$,

$$\lim_{n\to\infty} P\left(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \le x\right) = \Phi(x)$$

  where $\Phi(\cdot)$ is the normal c.d.f.
- saw graphs illustrating the DeMoivre-Laplace theorem for binomial random variables.

3 Sample Problems

Example 1 ***Spring 2003 Exam, Problem 3***
Mr Bayson, a third grade teacher in the Baldwin School in Cambridge is up for promotion, and the likelihood of it happening will depend in part on his students' performance on the MCAS exam. He has ten students and the exam will have ten questions on it. Suppose that each student has a 60% chance of correctly answering each questions, and that answers on all questions are independent. What is the probability that his top scoring student scores at least nine out of ten? What is the probability that his bottom-scoring student scores at least three out of ten?

Answer:
You should notice that this question has two parts: (1) determining the distribution of individual test scores, and (2) finding the c.d.f.s of the maximum and the minimum.
Since each student's exam score X is the number of successes out of 10 independent trials, it is a binomial random variable, $X \sim B(10, 0.6)$, for which we know the p.d.f.

$$f_X(x) = \begin{cases} \binom{10}{x} 0.6^x 0.4^{10-x} & \text{if } x \in \{0, 1, \ldots, 10\} \\ 0 & \text{otherwise} \end{cases}$$

In general, the maximum Y_1 of an i.i.d. sample $X_1, \ldots, X_n$ where the c.d.f. of X_i is $F_X(x)$ has c.d.f.

$$F_{Y_1}(y) = F_X(y)^n$$

and the minimum Y_2 has c.d.f.

$$F_{Y_2}(y) = 1 - (1 - F_X(y))^n$$

The probability that a *given* student scores less than 9 is

$$\begin{aligned} F_X(9) &= 1 - P(X = 9) - P(X = 10) = 1 - \binom{10}{9} 0.6^9 0.4 + \binom{10}{10} 0.6^{10} \\ &= 1 - 10 \cdot \frac{2 \cdot 3^9}{5^1 0} + \frac{3^{10}}{5^{10}} = 1 - \frac{23}{5} \cdot 0.6^9 \end{aligned}$$

Therefore, the probability that the top student scores at least 9 out of 10 is

$$1 - F_{Y_1}(9) = 1 - [F_X(9)]^{10} = 1 - \left(1 - \frac{23}{5} 0.6^9\right)^{10} \approx 37.79\%$$

The probability that a *given* student scores at least 3 out of 10 is

$$1-F_X(2) = 1-P(X=0)-P(X=1)-P(X=2) = 1-0.4^10-10\cdot 0.6\cdot 0.4^9-45\cdot 0.6^2\cdot 0.4^8 = 1-\frac{1876}{25}0.4^8$$

Therefore, the probability that the bottom student scores at least 3 out of ten is

$$1-F_Y(2) = [1-F_X(2)]^{10} \approx 60.39\%$$

Example 2 ***Spring 2007 Exam, Problem 3***
If $X \sim N(\mu, \sigma^2)$, we say that $Y = e^X$ has a lognormal distribution, $Y \sim L(\mu, \sigma^2)$.

(a) Find the p.d.f. of Y

(b) Suppose you have \$100,000 to invest and you have access to an investment whose return R_1 is distributed $L(\mu, \sigma^2)$. Its mean $e^{\mu+\sigma^2/2}$ is 1.10, and its variance $\left(e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}\right)$ is 0.01. What is the probability that your wealth at the end of one period of investment ($\$100,000R_1$) is greater than $110,000$?

(c) With the same parameter values as in (b), what is the probability that your wealth at the end of two independent periods of investment is greater than \$115,000?

Answer:

(a) This transformation is one-to-one, so we can use the change of variables formula. Note that X can be any real number, so the support of Y is $(0, \infty)$. The inverse transformation is $X = \ln(Y)$, which has $\frac{dX}{dY} = \frac{1}{X}$. Thus, using the change of variables formula, we have $f_Y(y) = \frac{1}{y}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(\frac{\ln(y)-\mu}{\sigma})^2)$ for $y > 0$ and $f_Y(y) = 0$ otherwise.

(b) It is useful to begin by solving for μ and σ^2. We can factor the expression for the variance as $e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$, or $[e^{\mu+\frac{1}{2}\sigma^2}]^2(e^{\sigma^2}-1)$. Plugging in from the expression for the mean and using the fact that the variance is .01, we have $(1.10)^2(e^{\sigma^2}-1) = .01$. Solving for σ yields $\sigma \approx .090722098$. We can then go back and see that $\mu \approx .09119493$.

Now let's find out the probability that your wealth at the end of one period is greater than \$110,000. We have $P(100000R_1 > 110000) = P(R_1 > 1.1) = P(\ln(R_1) > \ln(1.1)) = P(\frac{\ln(R_1)-\mu}{\sigma} > \frac{\ln(1.1)-\mu}{\sigma}) = 1-\Phi(\frac{\ln(1.1)-\mu}{\sigma}) \approx 1-\Phi(.045361051) \approx 1-.5181 = .4819$, where you can use the normal probability tables to get the value of the standard normal c.d.f.

(c) $P(100000R_1R_2 > 115000) = P(R_1R_2 > 1.15) = P(\ln(R_1)+\ln(R_2) > \ln(1.15))$. Note that $\ln(R_1)$ and $\ln(R_2)$ are independent normals, and so their sum is also normal. The mean is the sum of the means and the variance is the sum of the variances. Using tildes to refer to the new mean, variance, and standard deviation here, we thus have $\tilde{\mu} \approx .18238986$, $\tilde{\sigma}^2 \approx .016460998$, and $\tilde{\sigma} \approx .128300421$. So continuing along with the earlier calculation, we have $P(\ln(R_1)+\ln(R_2) > \ln(1.15)) = P(\frac{\ln(R_1)+\ln(R_2)-\tilde{\mu}}{\tilde{\sigma}} > \frac{\ln(1.15)-\tilde{\mu}}{\tilde{\sigma}}) = 1-\Phi(\frac{\ln(1.15)-\tilde{\mu}}{\tilde{\sigma}}) \approx 1-\Phi(-.3322508) \approx 1-.3698 = .6302$.

Example 3 ***Spring 2007 Exam, Problem 4***
Mikael Priks, a Swedish economist, has been studying various economic issues surrounding soccer hooliganism using detailed data on hooligan activity, fights, injuries, etc., collected by the Swedish police and self-reported by one of the gangs, the "Firman Boys" (see www.lrz-muenchen.de/ces/mikael.htm). In one paper he sought to analyze the determinants of the likelihood and severity of fights between rival hooligan

groups. To do so, he constructed a model of fights and injuries where the number of chance meetings between two rival groups in a season follows a $P(5)$ (Poisson with $\lambda = 5$) distribution. Furthermore, he assumed that at least one injury occurs in every fight and that, in fact, any number of injuries up to ten are all equally likely.

(a) Given those assumptions, what is the expected number of injuries that two rival groups inflict on each other in a given year? What is the variance of that quantity?

(b) Suppose instead that a fight only happened with probability one-half when a chance meeting of two rival groups occurred (you may assume independence of chance meetings). How would your answers to part (a) change?

Answer:

(a) Let us use X to refer to the number of fights in a season and Y to refer to the number of injuries. Here we'll assume that a chance meeting necessarily results in a fight. We have $E(Y) = E(E(Y|X)) = E(5.5X) = 5.5E(X) = 5.5(5) = 27.5$. And we have $Var(Y) = E(Var(Y|X)) + Var(E(Y|X)) = E(\frac{10^2-1}{12}X) + Var(5.5X)$. [Note that the variance of the number of injuries in a fight is $\frac{10^2-1}{12}$, and thus the variance of the number of injuries in X fights is $\frac{10^2-1}{12}X$ if the distribution of injuries is independent across fights.] Continuing along, we have $E(\frac{10^2-1}{12}X) + Var(5.5X) = \frac{99}{12}E(X) + \frac{121}{4}Var(X) = \frac{99}{12}(5) + \frac{121}{4}(5) = 192.5$.

(b) Let us use Z to refer to the number of chance encounters. We have $E(Y) = E(E(E(Y|X)|Z)) = E(E(5.5X|Z)) = E(5.5(E(X|Z))) = E(5.5\frac{1}{2}Z) = 2.75E(Z) = 2.75(5) = 13.75$. For the variance, we can still say that $Var(Y) = E(Var(Y|X)) + Var(E(Y|X)) = \frac{99}{12}E(X) + \frac{121}{4}Var(X)$, but now $E(X)$ and $Var(X)$ will have changed from part a. It is not hard to see that $E(X)$ is half of its previous value (so that now it is 2.5), and we can use the fact that $X|Z$ is a binomial with $p = .5$ and Z trials to write the variance of X as $Var(X) = E(Var(X|Z)) + Var(E(X|Z)) = E(Z(.5)(1-.5)) + Var(.5Z) = .25E(Z) + .25Var(Z) = .25(5) + .25(5) = 2.5$. So going back, we have $Var(Y) = \frac{99}{12}E(X) + \frac{121}{4}Var(X) = \frac{99}{12}(2.5) + \frac{121}{4}(2.5) = 96.25$.

Sample Problems

Spring 2003 Exam, Problem 3 Mr Bayson, a third grade teacher in the Baldwin School in Cambridge is up for promotion, and the likelihood of it happening will depend in part on his students' performance on the MCAS exam. He has ten students and the exam will have ten questions on it. Suppose that each student has a 60% chance of correctly answering each questions, and that answers on all questions are independent. What is the probability that his top scoring student scores at least nine out of ten? What is the probability that his bottom-scoring student scores at least three out of ten?

Spring 2007 Exam, Problem 3 If $X \sim N(\mu, \sigma^2)$, we say that $Y = e^X$ has a lognormal distribution, $Y \sim L(\mu, \sigma^2)$.

(a) Find the p.d.f. of Y

(b) Suppose you have \$100,000 to invest and you have access to an investment whose return R_1 is distributed $L(\mu, \sigma^2)$. Its mean $e^{\mu+\sigma^2/2}$ is 1.10, and its variance $\left(e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}\right)$ is 0.01. What is the probability that your wealth at the end of one period of investment ($\$100,000R_1$) is greater than $110,000$?

(c) With the same parameter values as in (b), what is the probability that your wealth at the end of two independent periods of investment is greater than \$115,000?

Spring 2007 Exam, Problem 4 Mikael Priks, a Swedish economist, has been studying various economic issues surrounding soccer hooliganism using detailed data on hooligan activity, fights, injuries, etc., collected by the Swedish police and self-reported by one of the gangs, the "Firman Boys" (see www.lrz-muenchen.de/ces/mikael.htm). In one paper he sought to analyze the determinants of the likelihood and severity of fights between rival hooligan groups. To do so, he constructed a model of fights and injuries where the number of chance meetings between two rival groups in a season follows a $P(5)$ (Poisson with $\lambda = 5$) distribution. Furthermore, he assumed that at least one injury occurs in every fight and that, in fact, any number of injuries up to ten are all equally likely.

(a) Given those assumptions, what is the expected number of injuries that two rival groups inflict on each other in a given year? What is the variance of that quantity?

(b) Suppose instead that a fight only happened with probability one-half when a chance meeting of two rival groups occurred (you may assume independence of chance meetings). How would your answers to part (a) change?