Section 2 Notes

Elizabeth Stone and Charles Wang

January 15, 2009

1 Joint, Marginal, and Conditional Probability

Useful Rules/Properties

1. $P(X=x) = \sum_i P(X=x, Y=y_i)$ or $\int y f_{XY}(x,y)\,dy$
2. $P(X=x|Y=y) = \frac{P(X=x,Y=y)}{P(Y=y)}$
3. $P(X=x, Y=y) = P(X=x|Y=y)\,P(Y=y)$
4. Bayes' Rule: $P(X=x|Y=y) = \frac{P(Y=y|X=x)P(X=x)}{\sum_i P(Y=y,X=x_i)}$

2 Expectation and Conditional Expectation of a Random Variable

2.1 Expectation

Definition 1 $E(X) = \begin{cases} \sum_i x_i p(X=x_i) & \text{if } X \text{ discrete RV} \\ \int x f_X(x)dx & \text{if } X \text{ continuous RV} \end{cases}$

Useful Properties

1. $E(aX+b) = aE(X) + b$ for a,b constant
2. $E(X+Y) = E(X) + E(Y)$
3. $E(h(X)+g(Y)) = E(h(X)) + E(g(Y))$ for arbitrary functions g and h

2.2 Conditional Expectation

Definition 2 $E(X|Y=y) = \begin{cases} \sum_i x_i p(X=x_i|Y=y) & \text{if } X \text{ discrete RV} \\ \int x f_{X|Y=y}(x)dx & \text{if } X \text{ continuous RV} \end{cases}$

Useful Rules/Properties

1. **A function of Y**
2. Take out what is being conditioned: $E(XY|X) = XE(Y|X)$
3. **Law of iterated expectations**: $E(X) = E[E(X|Y)]$

Proof.

$$\begin{aligned}
E(X) &= \int x f_x dx \\
&= \int_x x \left(\int_y f_{X,Y=y}\, dy \right) dx \qquad (2) \\
&= \int_x x \left(\int_y f_{X|Y=y} f_{Y=y} dy \right) dx \qquad (3) \\
&= \int_x \int_y x f_{X|Y=y} f_{Y=y}\, dy dx \qquad (4) \\
&= \int_y \int_x x f_{X|Y=y} f_{Y=y}\, dx dy \qquad (5) \\
&= \int_y \left(\underbrace{\int_x x f_{X|Y=y} dx}_{\equiv E(X|Y)} \right) f_{Y=y}\, dy \qquad (6) \\
&\equiv E(E(X|Y)) \qquad (7)
\end{aligned}$$

■

4. Usefulness of the Law of Iterated Expectations in 102B

Proposition 3 *If $E(\varepsilon|X) = 0$ then $Cov(\varepsilon, f(X)) = 0$ for any arbitrary function g*

Proof. *WTS: $E(\varepsilon f(X)) = 0$*

$$E(\varepsilon f(X)) = E[E(\varepsilon f(X)|X)] = E\left[f(X) \underbrace{E(\varepsilon|X)}_{=0} \right] = 0 \quad ■$$

Note: So, if we know $E(\varepsilon|X) = 0$, we can easily show that X and ε are uncorrelated. (Use the identity function)

3 Covariance, Variance, and Correlation

3.1 Covariance

Definition 4 *(Univariate) $Cov(X,Y) \equiv E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$ denoted by σ_{xy}*

(Vectors): $Cov(\mathbf{X}, \mathbf{Y}) \equiv E\left[(\mathbf{X} - E(\mathbf{X}))(\mathbf{Y} - E(\mathbf{Y}))^T \right] = E\left(\mathbf{X}\mathbf{Y}^T \right) - E(\mathbf{X})E(\mathbf{Y})^T$ denoted by Σ_{XX}

Useful Properties

1. $Cov(a + X, Y) = Cov(X, Y)$ for a constant
2. $Cov(aX, bY) = abCov(X, Y)$
3. $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$
4. $Cov(aW + bX, cY + dZ) = acCov(W, Y) + adCov(W, Z) + bcCov(X, Y) + bdCov(X, Z)$
5. $Var(X) = Cov(X, X)$
6. If $X \perp Y \Rightarrow Cov(X, Y) = 0$

3.2 Variance

Definition 5 *(Univariate)* $Var(X) \equiv E\left[(X - E(X))^2\right] = E\left(X^2\right) - E(X)^2$ *denoted by* σ_x^2

(Vectors): $Cov(\mathbf{X}, \mathbf{X}) \equiv E\left[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T\right] = E\left(\mathbf{X}\mathbf{X}^T\right) - E(\mathbf{X})E(\mathbf{X})^T$ *denoted by* Σ_{XX}

Useful Properties

1. $Var(X+Y) = Cov(X+Y, X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
2. $Var \geq 0$
3. Multivariate Example: Let $\mathbf{Z} = \begin{bmatrix} W \\ X \end{bmatrix}$.Then,

$$\begin{aligned} Var(\mathbf{Z}) &= E\left[(\mathbf{Z} - E(\mathbf{Z}))(\mathbf{Z} - E(\mathbf{X}))^T\right] \\ &= \begin{bmatrix} E\left[(W - EW)^2\right] & E\left[(W - EW)(X - EX)\right] \\ E\left[(W - EW)(X - EX)\right] & E\left[(X - EX)^2\right] \end{bmatrix} \\ &= \begin{bmatrix} \sigma_W^2 & \sigma_{WX} \\ \sigma_{WX} & \sigma_X^2 \end{bmatrix} \end{aligned}$$

3.3 Correlation

Definition 6 $Corr(X,Y) \equiv \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ *denoted by* ρ_{xy}

Useful Properties

1. $\rho_{xy} \in [-1, 1]$

   Proof. WTS: $\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \leq 1$ and $\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \geq -1$

   $\left(\text{same as showing } \left|\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}\right| \leq 1, \text{ or } Cov(X,Y) \leq \sqrt{Var(X)Var(Y)}\right)$

   Proof follows from Cauchy-Schwartz inequality. ■
2. $\rho_{xy} = 1$ or -1 iff $Y = a + bX$ (i.e. X and Y are linear transformations of each other)

4 (Weak) Law of Large Numbers and Central Limit Theorem

Proposition 7 *(LLN, Univariate) Let* $\{Z_1, ..., Z_n\}$ *be a sequence of independently and identically distributed (iid) random variables with* $E(Z_i) = \mu$ *and* $Var(Z_i) = \sigma^2$.

Then, $\overline{Z}_n \equiv \frac{1}{n}\sum_{i=1}^n z_i \to_P \mu$

(LLN, Vector) Let $\{\mathbf{Z}_1, ..., \mathbf{Z}_n\}$ *be a sequence of independently and identically distributed (iid) random vectors with* $E(\mathbf{Z}_i) = \boldsymbol{\mu}$ *and* $Var(\mathbf{Z}_i) = \Sigma \equiv E\left(\mathbf{Z}_i\mathbf{Z}_i^T\right) - E\mathbf{Z}_i E\mathbf{Z}_i^T$

Then, $\overline{\mathbf{Z}}_n \equiv \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i \to_P \boldsymbol{\mu}$

Proof. *One way to show that a random variable converges in probability to a constant is to show that Var of the random variable converges to 0 and the bias goes to 0*

Use Chebychev Indequality: $P\left(\left|\overline{Z}_n - \mu\right| \leq \varepsilon\right) \leq \frac{E\left[\left(\overline{Z}_n - \mu\right)^2\right]}{\varepsilon^2} = \frac{Var\left(\overline{Z}_n\right) + \left(E\left[\overline{Z}_n - \mu\right]\right)^2}{\varepsilon^2}$ *for any arbitrary* ε

$E\left(\overline{Z}_n\right) = \mu$, *and* $Var\left(\overline{Z}_n\right) = \frac{1}{n^2}\sigma^2 \to_{n\to\infty} 0$ *so that* $P\left(\left|\overline{Z}_n - \mu\right| \leq \varepsilon\right) \to_{n\to\infty} 0$ ■

Proposition 8 *(CLT, Univariate) Let* $\{Z_1, ..., Z_n\}$ *be iid with* $E(Z_i) = \mu$ *and* $Var(Z_i) = \sigma^2$. *Then,* $\frac{1}{\sqrt{n}}\sum_{i=1}^n z_i \to_D N\left(\mu, \sigma^2\right)$ *(or* $\sqrt{n}\left(\overline{Z}_n - \mu\right) \to_D N\left(0, \sigma^2\right)$ *)*

(CLT, Vector) Let $\{\mathbf{Z}_1, ..., \mathbf{Z}_n\}$ *be iid with* $E(\mathbf{Z}_i) = \boldsymbol{\mu}$ *and* $Var(\mathbf{Z}_i) = \Sigma \equiv E\left(\mathbf{Z}_i\mathbf{Z}_i^T\right) - E\mathbf{Z}_i E\mathbf{Z}_i^T$. *Then,* $\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i \to_D N(\boldsymbol{\mu}, \Sigma)$ *(or* $\sqrt{n}\left(\overline{\mathbf{Z}}_n - \boldsymbol{\mu}\right) \to_D N(0, \Sigma)$ *)*

5 Normal and Bivariate Normal Random Variables

Definition 9 *A bivariate normal random variable is a 2x1 vector of random variables, each being a normal random variable.*

e.g. If $X \overset{D}{=} N\left(\mu_X, \sigma_X^2\right), Y \overset{D}{=} N\left(\mu_Y, \sigma_Y^2\right)$ *,then* $\mathbf{Z} = \begin{bmatrix} X \\ Y \end{bmatrix} \overset{D}{=} N_2\left(\overrightarrow{\mu}, \mathbf{\Sigma}\right)$ *where* $\overrightarrow{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$ *and* $\mathbf{\Sigma} = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}$

Useful Properties

1. Linear transformations of normals are normally distributed: if $X \tilde{} N\left(\mu_X, \sigma_X^2\right)$, then for a, b constants, $aX + b \tilde{} N\left(a\mu_X + b, a^2\sigma_X^2\right)$
2. How to find the mean / variance of linear transformations of bivariate normals: If $\mathbf{Z} \overset{D}{=} N_2\left(\overrightarrow{\mu}, \mathbf{\Sigma}\right)$, then for A_{nx2} (matrix of constants) and b_{nx1} (vector of constants), $AZ + b \overset{D}{=} N_n\left(A\mu + b, A\Sigma A'\right)$
3. Conditional distribution: If $\begin{bmatrix} X \\ Y \end{bmatrix} \overset{D}{=} N_2\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}\right)$,

   then $Y|X = x \overset{D}{=} N\left(\mu_Y + \underbrace{\rho\frac{\sigma_Y}{\sigma_X}}_{\frac{\sigma_{XY}}{\sigma_X^2}}\left(x - \mu_X\right), \sigma_Y^2\left(1 - \rho^2\right)\right)$
4. Adding/Subtracting normal random variables results in a normal random variable: Let $X \overset{D}{=} N\left(\mu_X, \sigma_X^2\right), Y \overset{D}{=} N\left(\mu_Y, \sigma_Y^2\right)$ then $X + Y \overset{D}{=} N\left(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}\right)$

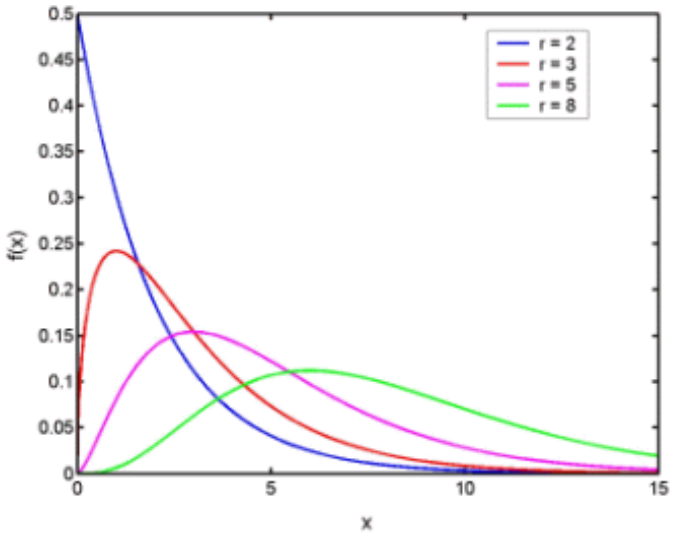
6 Chi-Squared Distribution

Definition 10 *Let* $Z_1, Z_2, ..., Z_p$ *be iid* $N(0,1)$*, then* $Z_1^2 + Z_2^2 + ... + Z_p^2 \overset{D}{=} \chi_p^2$

Proposition 11 *Suppose* $W_{k\times 1} \overset{D}{=} N\left(\mu_{k\times 1}, V\right)$*. Then* $T = \left(W - \mu\right)^T V^{-1}\left(W - \mu\right) \overset{D}{=} \chi^2(k)$

1. Clearly, $Z_1^2 \overset{D}{=} \chi_1^2$
2. Picture
3. Why is it useful in regressions?
   - Allows us to test hypothesis on multiple parameters at the same time (e.g. Suppose you have a regression model $Income = \beta_0 + \beta_1 Race + \beta_2 Gender + \varepsilon$ for which you have estimates $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right)$ for the population parameters. Suppose you want to test the hypothesis that race AND gender have no effect on income, i.e. $H_0 : \beta_1 = 0$ and $\beta_2 = 0$) –> USE WALD STATISTIC
   - Wald Statistic: We saw in class that under the null hypothesis that $\beta_{true} = b$ for some hypothesized β, the Wald Statistic $\sqrt{n}\left(\hat{\beta} - b\right)^T V^{-1}\sqrt{n}\left(\hat{\beta} - b\right)$ which is distributed $\chi^2(p)$.

     Note: $\sqrt{n}\left(\hat{\beta} - b\right)$ is normally distributed with mean 0 and variance V under the null.



- Example to show intuition behind why Wald statistic is chi-sq distributed:
  Let $X \overset{D}{=} \tilde{} N(0,1)$ and $Y \overset{D}{=} N(0,1)$ and X, Y independent. Let's write this in matrix notation:

$$\beta = \begin{bmatrix} X \\ Y \end{bmatrix} \tilde{} N \left(\underbrace{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}_{b}, \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{V=I_2} \right)$$

  Then,

$$\begin{aligned} T &= (\beta - b)^T V^{-1} (\beta - b) \\ &= \beta^T \beta \\ &= \begin{bmatrix} X & Y \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} \\ &= X^2 + Y^2 \overset{D}{=} \chi^2(2) \end{aligned}$$

  since X,Y standard normal and independent of each other.
- Another example: Suppose now $X \tilde{} N\left(0, \sigma_X^2\right)$ and $Y \tilde{} N\left(0, \sigma_Y^2\right)$ and X, Y independent
  Note that here $V = \begin{bmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{bmatrix}$ so that $V^{-1} = \begin{bmatrix} \frac{1}{\sigma_X^2} & 0 \\ 0 & \frac{1}{\sigma_Y^2} \end{bmatrix}$
  Then

$$\begin{aligned} T &= (\beta - b)^T V^{-1} (\beta - b) \\ &= \begin{bmatrix} X & Y \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_X^2} & 0 \\ 0 & \frac{1}{\sigma_Y^2} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} \\ &= \frac{X^2}{\sigma_X^2} + \frac{Y^2}{\sigma_Y^2} \\ &= \left(\underbrace{\frac{X}{\sigma_X}}_{N(0,1)} \right)^2 + \left(\underbrace{\frac{Y}{\sigma_Y}}_{N(0,1)} \right)^2 \overset{D}{=} \chi^2(2) \end{aligned}$$

since $\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}$ standard normal and independent of each other.

- What we do in the regression setting is essentially a more generalized version of this. (Don't need to know):

$$\begin{aligned} T &= (\beta - b)^T V^{-1} (\beta - b) \\ &= (\beta - b)^T C^T C (\beta - b) \quad \text{by Cholesky decomposition of } V^{-1} \\ &= [C(\beta - b)]^T [C(\beta - b)] \end{aligned}$$

The matrix C essentially re-weights the coordinates in such a way that elements of the new vector $C(\beta - b)$ are standard normals and are independent of each other, and therefore the inner dot product is just a sum of independent standard normal variables, which is chi-square distributed.

7 Extra Stuff (Don't Need to know)

7.1 Trilogy of Theorems (To help us figure out the limiting distribution for more complicated sequences of RVs)

1. Slutsky's Theorem: If $Y_n \to_D Y$ and $A_n \to_P a, B_n \to_P b$ for a, b non-random constants, then $A_n Y_n + B_n \to_D aY + b$
2. Continuous Mapping Theorem: If $Y_n \to_P Y$ and g is a continuous function, then $g(Y_n) \to_P g(Y)$
3. Delta Method (To be covered later)

An example using the above: Showing consistency of the OLS coefficient:
Suppose data is generated according to

$$Y_n = a + bx_n + \varepsilon$$

$$\text{where } E(\varepsilon|X) = 0 \quad (\Rightarrow Ex_i\varepsilon_i) = 0)$$

We have an iid sample of n datapoints, generated according to this process.
Question: What is the probability limit of the OLS coefficient and the limiting distribution of the OLS coefficient?

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1} X'Y \\ &= (X'X)^{-1} X'(X\beta + E) \\ &= \beta + (X'X)^{-1} X'E \\ &= \beta + \left(\frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \left(\frac{1}{n}\sum_i \mathbf{x}_i\varepsilon_i\right) \end{aligned}$$

By LLN, $\frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i' \to_P E(\mathbf{x}_i\mathbf{x}_i')$
By Continuous mapping, $\left(\frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \to_P [E(\mathbf{x}_i\mathbf{x}_i')]^{-1}$
By LLN, $\frac{1}{n}\sum_i \mathbf{x}_i\varepsilon_i \to_P E(\mathbf{x}_i\varepsilon_i) = 0$
So, by Slutsky's, $\hat{\beta} = \beta + \left(\frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \left(\frac{1}{n}\sum_i \mathbf{x}_i\varepsilon_i\right) \to_P \beta$