Dis 2: Stats Review; More on Stata

1 Stats Review

1.1 Random Variable and Sampling

- **Random variable**: Take some probabilistic event, assign the outcome of that event to a value.

  Ex. Let X be a random variable of playing rock-paper-scissor. Then one way the outcome of rock-paper-scissor can be recorded as a random variable is

$$\begin{aligned} &\text{Playing rock} && \rightarrow && X = 1 \\ &\text{Playing paper} && \rightarrow && X = 10 \\ &\text{Playing scissor} && \rightarrow && X = 234 \end{aligned}$$

  Ex. Let X be a random variable of a six-sided die rolling event. Then

$$\begin{aligned} &\text{Rolling a 1} && \rightarrow && X = 1 \\ &\text{Rolling a 2} && \rightarrow && X = 2 \\ &&& \cdots \\ &\text{Rolling a 6} && \rightarrow && X = 6 \end{aligned}$$

  Ex. Let X be a random variable of certain country's real GDP. Then X can be any real number.

- **Properties of common operations on random variables**

  Let X and Y be random variables, a, b, c, d be constants, $f(X)$ be some function applied onto X.

  - Mean (expected value)
    * $E(c) = c$
    * $E(aX + b) = aE(X) + b$
    * $E(X + Y) = E(X) + E(Y)$
    * $E[f(X)Y + c|X] = f(X)E[Y|X] + c$
    * Law of iterated expectation: $E[E(Y|X)] = E[Y]$
  - Variance
    * $Var(c) = 0$
    * $Var(aX + b) = a^2 Var(X)$
    * $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$
  - Covariance
    * $Cov(a, b) = 0$
    * $Cov(X, X) = Var(X)$
    * $Cov(aX + b, cY + d) = acCov(X, Y)$

- **Sample**: Among all the possible outcomes that a random variable can achieve, take a subset of it. This subset is called a sample, and we aim to use the sample to derive some statistics about the population.

Ex. Among all Dane county residents, take a sample of 100 people, and ask for their income level. The income of these 100 people is a sample of size 100, and it might be useful in telling us something about the income level in Dane county (such as the average / median / min / max).

- **Representative sample**: A sample constructed that exhibits characteristics typical of those possessed by the population.
- **Random sample**: A sample constructed so that the probability of selecting an observation in the sample is the same for all observations.

  A random sample is most likely representative.

  A representative sample though doesn't need to be randomly selected.

- **Population parameters vs. Sample estimators**
  - **Population parameters** (usually denoted in Greek letters): Have access to every single data point within the true population, and then perform some calculation to arrive at some statistics.

    Population parameters are relevant for random variables.
  - **Sample estimators** (usually denoted in Roman letters, or with a hat on top of the population parameter): Cannot access every data point within the population, but can extract a sample from the population. The sample is then used to perform calculation to arrive at some estimators that, ultimately, hope to approximate the corresponding statistics from the population.

    Sample estimators are relevant for the sample constructed from sampling outputs of a random variable.

	Population Parameter	**Sample Estimator**
Mean	$\mu = E[X] = \begin{cases} \sum_x xf(x) & \text{(discrete)} \\ \int xf(x)dx & \text{(continuous)} \end{cases}$	$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$
Variance	$\sigma^2 = Var(X) = E[(X-\mu)^2]$ $= E[X^2] - [E(X)]^2$	$s^2 = \widehat{Var}(X) = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$ $= \frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2$
Standard deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$
Covariance	$\sigma_{XY} = Cov(X,Y)$ $= E[(X-\mu_X)(Y-\mu_Y)]$ $= E(XY) - E(X)E(Y)$	$s_{XY} = \widehat{Cov}(X,Y)$ $= \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ $= \frac{1}{n}\sum_{i=1}^n x_i y_i - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n y_i\right)$
Coefficient of Correlation	$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$	$r = \hat{\rho} = \frac{s_{XY}}{s_X s_Y} = \frac{\widehat{Cov}(X,Y)}{\sqrt{\widehat{Var}(X)}\sqrt{\widehat{Var}(Y)}}$

1.2 Probability Theory

- **PDF / PMF and CDF**
  - For discrete random variable:
    - PMF (Probability Mass Function) measures point probability.

      Ex. For discrete random variable X, $f(1.3) = Pr(X = 1.3)$

    - CDF (Cumulative Distribution Function) measures probability up to certain point.
      Ex. For discrete random variable X, $F(1.3) = Pr(X \leq 1.3)$
  - For continuous random variables:
    - PDF (Probability Density Function) measures point density.
      Note that for continuous random variables, probability of hitting any point is 0, but its density at a point needs not to be 0.
    - CDF (Cumulative Distribution Function) measures density up to certain point.
      Ex. For continuous random variable X, $F(1.3) = \int_{-\infty}^{1.3} f(x)dx = Pr(X \leq 1.3)$

- **Joint, conditional, and marginal density / probability**

  Let X, Y be random variables, x, y be the value that X, Y can respectively take.

  - $f(x,y) = f_{XY}(x,y)$ is the density that $X = x$, $Y = y$ jointly occur.
    For discrete random variables, $f_{XY}(x,y) = Pr(X = x, Y = y)$.
    Ex. Let X be income, Y be years of schooling. $Pr(X = 100,000, Y = 16)$ is **among all population, what's the probability of someone who earns $100,000 and has 16 years of schooling at the same time**.
  - $f(x|y) = f_{X|Y}(x|y)$ is conditional on $Y = y$, what's the density of X at point x.
    For discrete random variables, $f_{X|Y}(x|y) = Pr(X = x|Y = y)$.
    Ex. Let X be income, Y be years of schooling. $Pr(X = 100,000|Y = 16)$ is **among people who has 16 years of schooling, what's the probability of these people earning an income of $100,000**.
  - $f_X(x)$ is the marginal density of X. Using a discrete example [*]:

	$X = 1$	$X = 2$	$X = 3$	$X = 4$	$f_Y(y)$
$Y = 1$	$4/32$	$2/32$	$1/32$	$1/32$	$8/32$
$Y = 2$	$3/32$	$6/32$	$3/32$	$3/32$	$15/32$
$Y = 3$	$9/32$	0	0	0	$9/32$
$f_X(x)$	$16/32$	$8/32$	$4/32$	$4/32$	

    To calculate marginal density of X:

$$f_X(x) = \begin{cases} \sum_{y \in \mathbb{Y}} f_{XY}(x,y) & \text{(discrete random variable)} \\ \int_{y \in \mathbb{Y}} f_{XY}(x,y)dy & \text{(continuous random variable)} \end{cases}$$

  - What's the relationship between the three?

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} \qquad Pr(X = x|Y = y) = \frac{Pr(X = x, Y = y)}{Pr(Y = y)}$$

    This also gives rise to **Bayes' Theorem**:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} \qquad Pr(X = x|Y = y) = \frac{Pr(Y = y|X = x)Pr(X = x)}{Pr(Y = y)}$$

[*] Example from Wikipedia

2 Problems: More on Stata

1. Let's take a closer look at the joint density table from the stats review part of this week's handout:

	$X=1$	$X=2$	$X=3$	$X=4$
$Y=1$	$4/32$	$2/32$	$1/32$	$1/32$
$Y=2$	$3/32$	$6/32$	$3/32$	$3/32$
$Y=3$	$9/32$	0	0	0

(a) Write down the expression for calculating the marginal distribution of X (i.e. $f_X(x)$).

The marginal distribution of X is

$$f_X(x) = \sum_{y=1}^{3} f_{XY}(x,y)$$

For example, when $X=1$, then

$$\begin{aligned} f_X(x=1) &= \sum_{y=1}^{3} f_{XY}(x=1,y) \\ &= f_{XY}(x=1,y=1) + f_{XY}(x=1,y=2) + f_{XY}(x=1,y=3) \\ &= 4/32 + 3/32 + 9/32 = 16/32 \end{aligned}$$

and the same exercise can be applied to $x = 2,3,4$.

(b) What's the expression for conditional expectation of Y given X (i.e. what is $E[Y|X]$)?

The conditional expectation of Y given X is

$$E[Y|X] = \sum_{y=1}^{3} y * f_{Y|X}(y|x)$$

For example, when $X=4$,

$$\begin{aligned} E[Y|X=4] &= \sum_{y=1}^{3} y * f_{Y|X}(y|x=4) \\ &= \sum_{y=1}^{3} y * \frac{f_{XY}(x=4,y)}{f_X(x=4)} \\ &= 1 * \frac{f_{XY}(x=4,y=1)}{f_X(x=4)} + 2 * \frac{f_{XY}(x=4,y=2)}{f_X(x=4)} + 3 * \frac{f_{XY}(x=4,y=3)}{f_X(x=4)} \\ &= 1 * \frac{1/32}{4/32} + 2 * \frac{3/32}{4/32} + 3 * \frac{0}{4/32} = 1.75 \end{aligned}$$

and the same can be done for $X = 1,2,3$.

(c) Write down the expression for the unconditional expectation of Y (i.e. $E[Y]$), using law of iterated expectation.

Using law of iterated expectation,

$$\begin{aligned} E[Y] &= E[E(Y|X)] \\ &= \sum_{x=1}^{4} E(Y|X=x) f_X(x) \end{aligned}$$

(d) Another way to calculate the unconditional expectation of Y is to use the marginal distribution of Y (i.e. $f_Y(y)$). Write down the expression for using this method.

Using marginal distribution, $E[Y]$ can be directly found from its definition

$$\begin{aligned} E[Y] &= \sum_{y=1}^{3} y * f_Y(y) \\ &= 1 * f_Y(1) + 2 * f_Y(2) + 3 * f_Y(3) \end{aligned}$$

$f_Y(y)$ can be found using the same technique in part (a) of this question.

(e) What's the expression for calculating the variance of Y?

Variance of Y is

$$Var(Y) = E(Y^2) - [E(Y)]^2$$

Given that we have $E(Y)$ expressed in the earlier part of this question, we only need the express $E(Y^2)$ now. Using marginal distribution,

$$\begin{aligned} E(Y^2) &= \sum_{y=1}^{3} y^2 * f_Y(y) \\ &= 1^2 * f_Y(1) + 2^2 * f_Y(2) + 3^2 * f_Y(3) \end{aligned}$$

2. Load this week's dataset into Stata (don't forget to first change your working directory). This week's dataset is exactly the joint distribution given in problem 1. All Stata commands should be written in a do-file.

   (a) With the dataset we have, can we calculate the expectation of Y by simply running something like `mean(Y)`? If not, why?

   We cannot directly calculate the expectation of Y by running something like `mean(Y)`. This is because the dataset we have describes the **joint density between X and Y, instead of the actual data points (X, Y)**.

   If we have data like the following:

$$
\begin{gathered}
(X_1, Y_1) \\
(X_2, Y_2) \\
\dots \\
(X_n, Y_n)
\end{gathered}
$$

   then we know for each of the n observations, what is the value of X and Y there, and this would allow us to calculate the mean of Y by using something like `mean(Y)`.

   (b) Calculate the conditional expectation of Y given X.

   With reference to the expression given in question 1, see this week's do-file solution.

   (c) Calculate the unconditional expectation of Y, using law of iterated expectation.

   With reference to the expression given in question 1, see this week's do-file solution.

   (d) Calculate the unconditional expectation of Y, using the marginal distribution of Y.

   With reference to the expression given in question 1, see this week's do-file solution.

   (e) Calculate the variance of Y.

   With reference to the expression given in question 1, see this week's do-file solution.

   (f) Save the results from running your do-file into a log.

   See this week's do-file solution.

3. (Part of problem #5 from this week's problem set)

   The formula for the normal distribution is

$$f(Y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-\frac{1}{2}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2}$$

   SAT scores in Mathematics are normally distributed with a mean of 500 and a standard deviation of **100** ← (did not get printed on your problem set). Create the following dataset in Stata:

   (a) In the first column, start by entering 300 (the lowest score on the mathematics section), and then increment the scores by 10 until you reach 800.

   See this week's do-file solution.

   (b) In the second column, use the formula for the normal distribution and calculate $f(\text{math\_SAT})$

   Notice that SAT scores in math (math_SAT) is normally distributed with a mean of 500 and a standard deviation of 100, which means from the normal density formula,

   - Y random variable should record the possible math SAT scores
   - $\mu_Y = 500$
   - $\sigma_Y = 100$

   We can then use the formula for density provided in the question to calculate $f(\text{math\_SAT})$ (see this week's do-file for specific commands).

   Side note #1: Another way to accomplish this is by using the `normalden` function in Stata. Use the command

```
help normalden
```

   to understand its syntax.

   Side note #2: Notice that when we say math_SAT is normally distributed with a mean of 500 and a standard deviation of 100, it means that

$$\text{math\_SAT} \sim N(\mu, \sigma^2) = N(500, 100^2)$$

   The highlight in here is that the second argument in the $N(\mu, \sigma^2)$ notation records **variance**, which is **squared standard deviation**. This comes up again in some other part of your problem set this week.