Recitation 2

Sean Hyland

UN3412 Introduction to Econometrics

Spring 2020

Q3 : Functions of Random Variables

Consider the following table:

	Rain (X=0)	No Rain (X=1)	Total
Long Commute (Y=0)	0.15	0.07	0.22
Short Commute (Y=1)	0.15	0.63	0.78
Total	0.30	.70	1.00

Define the following random variables:

- $W = 3 + 6X$
- $V = 20 - 7Y$

Compute (a) $E(W)$ and $E(V)$

(b) $Var(W)$ and $Var(V)$

(c) $Cov(W,V)$ and $Corr(W,V)$.

Before proceeding, note the following:

- $E(X) = \sum_x x \cdot P(X = x) = P(X = 1) = 0.7$

  where the second equality follows because X is a Bernoulli random variable.

  Similarly, $E(Y) = P(Y = 1) = 0.78$.

- $Var(X) = \sum_{x=0}^{1} P(X = x) \cdot (x - E(X))^2$

  $= P(X = 0) \cdot (0 - E(X))^2 + P(X = 1) \cdot (1 - E(X))^2$

  $= (1 - P(X = 1)) \cdot (-P(X = 1))^2 + P(X = 1)(1 - P(X = 1))^2$

  $= (1 - P(X = 1)) \cdot P(X = 1) = 0.3 \times 0.7 = 0.21$

Similarly, $Var(Y) = (1 - P(Y = 1)) \cdot P(Y = 1) = 0.1716$

- $Cov(X,Y) = \sum_{x=0}^{1}\sum_{y=0}^{1} P(X=x, Y=y)(x-E(X))(y-E(Y))$

$= 0.15 \times (0-0.7)(0-0.78)$

$+ 0.15 \times (0-0.7)(1-0.78)$

$+ 0.07 \times (1-0.7)(0-0.78)$

$+ 0.63 \times (1-0.7)(1-0.78)$

$= 0.15 \times 0.546 + 0.15 \times -0.154$

$+ 0.07 \times -0.234 + 0.63 \times 0.066$

$= 0.084$

(a) Compute $E(w)$ and $E(v)$.

Consider the more general problem:
Let Z be a discrete random variable, and a, b any (real-valued) constants. Then

$$\begin{aligned} E[a+bZ] &= \sum_z P(Z=z)(a+bz) \\ &= a \cdot \sum_z P(Z=z) + b \cdot \sum P(Z=z) \cdot z \\ &= a \cdot 1 + b \cdot E[Z] \\ &= a + b\,E[Z] \end{aligned}$$

Note $\sum_z P(Z=z) = 1$.

$$\Rightarrow E[w] = 3 + 6E[X] = 3 + 6 \times 0.7 = 7.2$$
$$E[v] = 20 - 7E[Y] = 20 - 7 \times 0.78 = 14.54$$

(b). Compute $Var(W)$, $Var(V)$.

Let Z be a discrete random variable, and a, b any (real-valued) constants. Then

$$\begin{aligned} Var(a+bZ) &= E\left[(a+bZ-E(a+bZ])^2\right] \\ &= E[(a+bZ-a+bE[Z])^2] \\ &= E[(bZ-bE[Z])^2] \\ &= b^2 E[(Z-E[Z])^2] \\ &= b^2 Var(Z) \end{aligned}$$

$$\Rightarrow Var(W) = Var(3+6X) = 6^2 \cdot Var(X) = 7.56$$
$$Var(Y) = Var(20-7Y) = (-7)^2 \cdot Var(Y) = 8.41$$

(c) Compute $Cov(W,V)$ and $Corr(W,V)$

$$
\begin{aligned}
Cov(W,V) &= E[(W - E[W])(V - E[V])] \\
&= E[((3+6X)-(3+6E[X]))((20-7Y)-(20-7E[Y]))] \\
&= E[(6X-6E[X])(-7Y+7E[Y])] \\
&= -6\times 7 \cdot E[(X-E[X])(Y-E[Y])] \\
&= -42 \times Cov(X,Y) \\
&= -42 \times 0.084 = -3.528
\end{aligned}
$$

$$
\begin{aligned}
Corr(W,V) &= Cov(W,V) / (SD(W) \cdot SD(V)) \\
&= -3.528 / \sqrt{7.56 \times 8.4084} \\
&= -0.4425 \quad (4dp).
\end{aligned}
$$

Q4 : Employment & Education.

The following table gives the joint probability distribution between employment status and college graduation among those either employed or looking for work (unemployed) in the working age US population, based on the 1990 US Census.

	Unemployed (Y=0)	Employed (Y=1)	Total
Non-college grads (X=0)	0.045	0.709	0.754
College grads (X=1)	0.005	0.241	0.246
Total	0.050	0.950	1.000

(a) Compute $E(Y)$.

$E(Y) = P(Y=1) = 0.95$. Because Y is a Bernoulli R.V.

(b) Show $P(Y=0) = 1 - E(Y)$.

$$P(Y=0) + P(Y=1) = 1$$
$$\Rightarrow P(Y=0) = 1 - P(Y=1) = 1 - E(Y).$$

(c) Calculate $E(Y|X=1)$, $E(Y|X=0)$.

$$E(Y|X=1) = 0 \cdot P(Y=0|X=1) + 1 \cdot P(Y=1|X=1)$$
$$= P(Y=1|X=1)$$
$$= P(Y=1, X=1) / P(X=1) \quad \text{Bayes rule.}$$
$$= 0.241 / 0.246 = 0.9797.$$

$$E(Y|X=0) = P(Y=1|X=0) \quad \text{Bayes rule.}$$
$$= P(Y=1, X=0) / P(X=0)$$
$$= 0.709/0.754 = 0.9403.$$

(d) Calculate the unemployment rate for (i) college grads and (ii) non-college grads.

$$P(Y=0|X=1) = 1 - P(Y=1|X=1)$$
$$= 1 - E(Y|X=1) = 0.0203$$

$$P(Y=0|X=0) = 1 - P(Y=1|X=0)$$
$$= 1 - E(Y|X=0) = 0.0597$$

(c) What is the probability that a randomly selected unemployed worker is (i) a college graduate? (ii) a non-college graduate?

$$P(X=1 \mid Y=0) = P(X=1, Y=0) / P(Y=0)$$
$$= 0.005 / 0.05 = 0.1$$
$$P(X=0 \mid Y=0) = 1 - P(X=1 \mid Y=0) = 0.9$$

(f) Are X and Y independent?

No. Two random variables are indep. if $P(X=x, Y=y) = P(X=x) \cdot P(Y=y) \; \forall x, y$. But this implies $P(Y=y \mid X=x) = P(Y=y)$, however, in our example, the employment rate varies with education i.e. $P(Y=1 \mid X=0) \neq P(Y=1 \mid X=1)$, which violates independence.

Q5 : Central Limit Theorem (i)

Suppose $\mu_Y = 100$, $\sigma_Y^2 = 43$.
Calculate the following.

(a) $P(\bar{Y} \leq 101)$, with $N = 100$

We know the asymptotic distribution of $\bar{Y}$ is $N(\mu_Y, \sigma^2/N)$. Therefore

$$P(\bar{Y} \leq 101) = P\left(\frac{\bar{Y} - \mu_Y}{SD(\bar{Y})} \leq \frac{101 - \mu_Y}{SD(\bar{Y})}\right)$$
$$= P\left(\frac{\bar{Y} - \mu_Y}{\sqrt{\sigma_Y^2/N}} \leq \frac{101 - \mu_Y}{\sqrt{\sigma_Y^2/N}}\right)$$
$$= P\left(Z \leq \frac{101 - 100}{\sqrt{43/100}}\right)$$
$$= P(Z \leq 1.525) \approx 0.9364$$

(b) $Pr(\bar{Y} > 98)$, with $N = 165$

$$Pr(\bar{Y} > 98) = P\left(\frac{\bar{Y} - \mu_Y}{SD(\bar{Y})} > \frac{98 - \mu_Y}{SD(\bar{Y})}\right)$$

$$= P\left(\frac{\bar{Y} - \mu_Y}{\sigma_Y/\sqrt{N}} > \frac{98 - \mu_Y}{\sigma_Y/\sqrt{N}}\right)$$

$$= P\left(Z > \frac{98 - 100}{\sqrt{43}/165}\right)$$

$$= P(Z > -3.9178)$$

$$\approx 1$$

(c). $P(101 \leq \bar{Y} \leq 103)$, with $N = 64$.

$$P(101 \leq \bar{Y} \leq 103) = P(\bar{Y} \leq 103) - P(\bar{Y} < 101)$$

$$= P\left(\frac{\bar{Y} - \mu_Y}{\sqrt{\sigma_Y^2/N}} \leq \frac{103 - \mu_Y}{\sqrt{\sigma_Y^2/N}}\right) - P\left(\frac{\bar{Y} - \mu_Y}{\sqrt{\sigma_Y^2/N}} < \frac{101 - \mu_Y}{\sqrt{\sigma_Y^2/N}}\right)$$

$$= P\left(Z \leq \frac{103 - 100}{\sqrt{43/64}}\right) - P\left(Z < \frac{101 - 100}{\sqrt{43/64}}\right)$$

$$= P(Z \leq 3.6599) - P(Z < 1.22)$$

$$= 0.9999 - 0.8888$$

$$= 0.1111$$

Q6: Gender & Earnings.

To investigate possible gender discrimination in a firm, a sample of 100 men and 64 women with similar job descriptions are selected at random. A summary of the resulting monthly salaries are:

	Avg. Salary ($\bar{Y}$)	Stand Dev (of Y)	n
Men	\$3100	\$200	100
Women	\$2900	\$320	64

(a) What do these data suggest about wage differences in the firm? Do they represent statistically significant evidence that wages of men and women are different? (To answer this question, first state the null and alternative hypothesis; second, compute the relevant t-statistic; and finally, use the p-value to answer the equation.)

We first note that average salary among males in the sample is $200 greater than that of females – but is this a statistically significant difference?

$H_0: \mu_M = \mu_F$; $H_a: \mu_M \neq \mu_F$.

$$SE(\bar{Y}_M - \bar{Y}_F) = \sqrt{\frac{S_M^2}{n_M} + \frac{S_F^2}{n_F}} = \sqrt{\frac{200^2}{100} + \frac{320^2}{64}} = 44.72$$

$$\Rightarrow t^{act} = (\bar{Y}_M - \bar{Y}_F) / SE(\bar{Y}_M - \bar{Y}_F) = 200/44.72 = 4.472.$$

$$p\text{-value} = P(t > |t^{act}|) = 2 \cdot P(Z < -4.472) \approx 0$$

Thus the difference is statistically significant and we reject the null hypothesis

(b) Do these data suggest that the firm is guilty of gender discrimination in its compensation politics? Explain.

Part (a) shows a statistically significant difference in earnings; to test whether women earn less we should use a one-tailed test, which would suggest women do earn less than men on average. The results here are consistent with, but not necessarily proof of, gender discrimination. To prove the latter we would want to compare male and female workers who are otherwise identical.

7. [Practice question, not graded] SW 2.10 [Hint: Use SW Appendix Table 1.]

Compute the following probabilities:

(a) If Y is distributed N(1,4), find Pr(Y≤3).

(b) If Y is distributed N(3,9), find Pr(Y>0).

(c) If Y is distributed N(50,25), find Pr(40≤Y≤52).

(d) If Y is distributed N(5,2), find Pr(6≤Y≤8)

(a) $P(Y \le 3) = P\left(\frac{Y-1}{\sqrt{4}} \le \frac{3-1}{\sqrt{4}}\right)$

$= P(Z \le 1) = 0.8413.$

(b) $P(Y > 0) = 1 - P\left(\frac{Y-3}{\sqrt{9}} \le \frac{0-3}{\sqrt{9}}\right)$

$= 1 - P(Z \le -1) = P(Z \le 1) = 0.8413.$

(c) $P(40 \leq Y \leq 52) = P(Y \leq 52) - P(Y < 40)$

$= P\left(\frac{Y-50}{\sqrt{25}} \leq \frac{52-50}{\sqrt{25}}\right) - P\left(\frac{Y-50}{\sqrt{25}} < \frac{40-50}{\sqrt{25}}\right)$

$= P(Z \leq 0.4) - P(Z < -2)$

$= P(Z \leq 0.4) - (1 - P(Z < 2))$

$= 0.6554 - 1 + 0.9772$

$= 0.6326.$

(d) $P(6 \leq Y \leq 8) = P(Y \leq 8) - P(Y < 6)$

$= P\left(\frac{Y-5}{\sqrt{2}} \leq \frac{6-5}{\sqrt{2}}\right) - P\left(\frac{Y-5}{\sqrt{2}} < \frac{6-5}{\sqrt{2}}\right)$

$= P\left(Z \leq \frac{3}{\sqrt{2}}\right) - P\left(Z < \frac{1}{\sqrt{2}}\right)$

$= P(Z \leq 2.1213) - P(Z < 0.7071)$

$= 0.9831 - 0.7602 = 0.2229.$

8. [Practice question, not graded] SW 3.3

In a survey of 400 likely voters, 215 responded that they would vote for the incumbent and 185 responded that they would vote for the challenger. Let p denote the fraction of all likely voters that preferred the incumbent at the time of the survey, and let $\hat{p}$ be the fraction of survey respondents that preferred the incumbent.

(a) Use the survey results to estimate p.

(b) Use the estimator of the variance of $\hat{p}$, $\hat{p}\ (1 - \hat{p})/n$ to calculate the standard error of your estimator.

(c) What is the p-value for the test H0: p=0.5 vs. H1:p≠0.5?

(d) What is the p-value for the test H0: p=0.5 vs. H1:p>0.5?

(e) Why do the results from (c) and (d) differ?

(f) Did the survey contain statistically significant evidence that the incumbent was ahead of the challenger at the time of the survey? Explain.

(a) $\hat{p} = 215/400 = 0.5375$.

(b) $Var(\hat{p}) = \hat{p}(1-\hat{p})/n = 6.1248 \times 10^{-4}$

$\Rightarrow SE(\hat{p}) = \sqrt{Var(\hat{p})} = 0.0249$.

(c) $t^{act} = \frac{\hat{p} - p_0}{SE(\hat{p})} = \frac{0.5375 - 0.5}{0.0249} = 1.506$.

$P(t > |t^{act}|) = 2 \cdot P(Z < -1.506) = 0.132$.

↑ use standard normal distribution.

← as $N = 400$.

(d) $P(t > t^{act}) = 1 - P(Z < 1.506) = 0.066$

(e) One-tailed vs. two-tailed tests.

(f) Fail to reject the null of equality-

9. Consider two events A and B with Pr(A) = 0.5 and Pr(B) = 0.9. Determine the maximum and minimum values of Pr(A ∪ B).

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 0.5 + 0.9 - P(A \cap B).$$

$P(A) + P(B) = 1.4$, but probabilities cannot exceed one, so $P(A \cup B)$ can be at most equal to one.

$P(A \cup B)$ is minimized when $P(A \cap B)$ is minimized, which occurs when $A \subseteq B$.

In this case $P(A \cap B) = P(A) = 0.5$, so $P(A \cup B)$ must be at least 0.9.

10. Assume that events A and B^c are independent. That is, $\Pr(A \cap B^c) = \Pr(A)\Pr(B^c)$. Are events A and B also independent?

$$P(A) = P(A \cap B^c) + P(A \cap B)$$

$$\Rightarrow P(A \cap B) = P(A) - P(A \cap B^c)$$

$$= P(A) - P(A) \cdot P(B^c)$$

$$= P(A)(1 - P(B^c))$$

$$= P(A) \cdot P(B).$$

Therefore if A and B^c are independent, it follows that A and B must be independent.

Q11.

(i) Show $Cov(X,Y) = E[XY]$ if $E[X]=0$ or $E[Y]=0$.

$$Cov(X,Y) = E[XY] - E[X]\cdot E[Y]$$
$$= E[XY] \quad \text{if} \quad E[X]=0 \text{ or } E[Y]=0.$$

(a) Show $Cov(a+\alpha X, b+\beta Y) = \alpha\beta Cov(X,Y)$ for constants a, α, b, β.

$$Cov(a+\alpha X, b+\beta Y)$$
$$= E[(a+\alpha X - E[a+\alpha X])(b+\beta Y - E[b+\beta Y])]$$
$$= E[(\alpha X - \alpha\mu_X)(\beta Y - \beta\mu_Y)]$$
$$= \alpha\beta E[(X-\mu_X)(Y-\mu_Y)] = \alpha\beta Cov(X,Y)$$

(b) Show $Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y)$

$$
\begin{aligned}
Var(X+Y) &= E[(X+Y)^2] - (E[X+Y])^2 \\
&= E[X^2 + Y^2 + 2XY] - (\mu_X^2 + \mu_Y^2 - 2\mu_X\mu_Y) \\
&= (E[X^2] - \mu_X^2) + (E[Y^2] - \mu_Y^2) + (2E[XY] - 2\mu_X\mu_Y) \\
&= Var(X) + Var(Y) + 2Cov(X,Y)
\end{aligned}
$$

(c) Show $Corr(X,Y) \in [-1,1]$. $\forall\ X, Y$.

Cauchy-Schwarz: $Cov^2(X,Y) \leq Var(X) \cdot Var(Y)$

$$
\begin{aligned}
&\Rightarrow -SD(X)\cdot SD(Y) \leq Cov(X,Y) \leq SD(X)\cdot SD(Y) \\
&\Rightarrow -1 \leq \frac{Cov(X,Y)}{SD(X)\cdot SD(Y)} \leq 1.
\end{aligned}
$$

12. The following admission data are for the graduate program in the six largest majors at the University of California at Berkeley for the fall 1973 quarter.

Graduate	Male Applicants		Female Applicants	
Program	Admitted	Rejected	Admitted	Rejected
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317

(a) What is the overall probability of being admitted for males? For females? What is the standard deviation for males and for females?

Number of male admittees: $512 + \cdots + 22 = 1198$.

Number of male rejectees: $313 + \cdots + 351 = 1493$

$$\Rightarrow \hat{p}_m = \frac{1198}{1198+1493} = 0.4452$$

$$SD(\hat{p}_m) = \sqrt{\hat{p}_m(1-\hat{p}_m)} = 0.4970$$

Note: SD of outcome, not SE of estimator.

Number of female admittees: $89 + \cdots + 24 = 557$

Number of female rejectees: $19 + \cdots + 317 = 1278$

$$\Rightarrow \hat{p}_m = 557/557 + 1278 = 0.3035$$

$$SD(\hat{p}_m) = \sqrt{\hat{p}_m(1-\hat{p}_m)} = 0.4598$$

(b) How would you write down the null and alternative hypotheses in order to test that the overall probability of admission is higher for men than for women?

$H_0: p_m \leq p_F$; $H_a: p_m > p_F$

(c) Conduct a t-test of the hypothesis from part (b) and report the p-value.

Let us assume that, even if the means are different, the variances are the same.

$\Rightarrow$ Pooled variance t-test.

$$S_p^2 = \frac{(n_M - 1)\cdot S_M^2 + (n_F - 1)\cdot S_F^2}{n_M + n_F - 2}$$

$$= \frac{2690 \times 0.2470 + 1835 \times 0.2114}{2690 + 1835}$$

$$= 664.4182 + 387.7156 / 4526$$

$$= 0.2325$$

$$\Rightarrow SE_{\hat{p}}(\hat{p}) = S_p \cdot \sqrt{\frac{1}{n_M} + \frac{1}{n_F}} = 0.4821 \times \sqrt{\frac{1}{2691} + \frac{1}{1836}} = 0.0146$$

$$\Rightarrow t^{act} = \hat{p}_M - \hat{p}_F / SE(\hat{p}) = 0.1416 / 0.0146 = 9.7038$$

p-value $= P(Z > 9.7038) < 0.0001$.

(d) Is the result significant at the 5% level? Does it provide evidence of discrimination?

Yes, the pvalue is less than 0.05, so we would reject the null hypothesis at the 5% level.
This is consistent with, but may not be caused by, discrimination. Further investigation is warranted.

(e) Committee chairpersons claim they are more likely to admit women than men. Is this claim true? Compute acceptance rates for men and women by graduate program.

Repeating the process for each individual program reveals that females had higher acceptance rates than males in programs A, B, D and F, but lower acceptances in programs C and E. (Note the difference in acceptance rates is highly significant for program A, but not statistically significant in any other program). The overall result then is driven by female applications being concentrated in programs with lower acceptance rates.

(e) Committee chairpersons claim they are more likely to admit women than men. Is this claim true? Compute acceptance rates for men and women by graduate program.

See table below for the program-specific analysis.

Graduate	Male Applicants				Female Applicants				Statistics			
Program	Admitted	Rejected	% Admitted	Variance	Admitted	Rejected	% Admitted	Variance	Difference	Pooled Var.	Std. Error	t-statistic
A	512	313	0.6206	0.2355	89	19	0.8241	0.1450	-0.2035	0.2251	0.0485	-4.1913
B	353	207	0.6304	0.2330	17	8	0.6800	0.2176	-0.0496	0.2324	0.0985	-0.5038
C	120	205	0.3692	0.2329	202	391	0.3406	0.2246	0.0286	0.2275	0.0329	0.8684
D	138	279	0.3309	0.2214	131	244	0.3493	0.2273	-0.0184	0.2242	0.0337	-0.5460
E	53	138	0.2775	0.2005	94	299	0.2392	0.1820	0.0383	0.1880	0.0382	1.0014
F	22	351	0.0590	0.0555	24	317	0.0704	0.0654	-0.0114	0.0602	0.0184	-0.6199
All	1198	1493	0.4452	0.2470	557	1278	0.3035	0.2114	0.1416	0.2326	0.0146	9.7017

This shows females were more likely to be admitted to programs A, B, D and F than male counterparts (although the only difference which is statistically significant is that of A

Note: In no program are females significantly less likely to be admitted.

However, 50% of male applications are made to programs A or B, which have an overall admit rate > 0.6, whereas just 5% of female applications are submitted to those programs.

The overall result then is driven by female applications being concentrated in programs with lower acceptance rates.