QE sheet 2

(note, Q worded a bit like it wants you to find the cdf...)

4a) We're interested in the distribution of N, where N is the number of trials such that $\sum_{i=1}^{N} X_i = 1$, with each X_i iid $\sim$ Bernoulli(p).

$N = 1$ w.p. p ie success on trial 1
$N = 2$ w.p. $(1-p)p$ ie XV, and so on.
So $P(N = n) = (1-p)^{n-1} \cdot p$ for all $n \in \mathbb{N}^+$ ✓

~~Actually I think I misread Q, we actually want~~ $(1-p)^n$

b) $P(N=n) = \begin{cases} 0 & \text{if } n = 1 \\ 1 - p^n - (1-p)^n \; ? & \text{otherwise} \end{cases}$

I don't follow...

for $n \in \mathbb{N}^+$

For $n \geq 2$, in order to not meet the condition, we require either all ✓s or all Xs

7a)i f_X: $P(X=0) = 0.3$, $P(X=1) = 0.7$ ✓ $\quad f_Y$: $P(Y=0) = 0.4$, $P(Y=1) = 0.1$, $P(Y=2) = 0.5$ ✓

Marginal pmf is pmf regardless of others' values

ii.

x \ y	0	1	2
0	$1/3$	0	$2/3$
1	$3/7$	$1/7$	$3/7$

numbers in cells are $P(Y=y \mid X=x)$ ✓

b). Not independent. $X \perp\!\!\!\perp Y \iff$ the distribution of Y does not depend on the value of X, i.e. $f_{Y|X=x}$ is the same for all x. Clearly this isn't true, as shown above. ✓ ↑ (or $P(Y|X=x) = P(Y)$, ie different notation!)

c) $\mathbb{E}[Y] = 0 \times 0.4 + 1 \times 0.1 + 2 \times 0.5 = 1.1$ ✓
$\mathbb{E}[Y|X=1] = 1/7 + 2 \times 3/7 = 1$ ✓
$\mathbb{E}[Y|X=0] = 4/3$ ✓

and $0.3 \times 4/3 + 0.7 \times 1 = 1.1$ as expected. = Hah.

d) $\mathbb{E}[Y|X] \stackrel{\text{def}}{=} \sum_{y \in \{0,1,2,3\}} y \cdot P(Y=y|X)$

Yes [This object is a function of the random variable X, and thus itself a random variable.

9a) $\text{Cov}(Z, aX+bY+c) \stackrel{\text{def}}{=} E[Z \cdot (aX+bY+c)] - E[Z]\,E[aX+bY+c]$ ✓

$= E[Z \cdot aX] - E[Z]E[aX] + E[Z \cdot bY] - E[Z]E[bY] + E[Z \cdot c] - E[Z] \cdot E[c]$ by linearity of E

$= a \cdot (E[Z \cdot X] - E[Z]E[X]) + b(\cdots) + c \cdot (E(Z) - E(Z))$

by properties of E for constants, scaling, also from linearity

$\stackrel{\text{def}}{=} a\,\text{Cov}(Z,X) + b\,\text{Cov}(Z,Y)$ as required ✓

b) As noted, let $Z := aX + bY + c$.

$\text{Var}(Z) \stackrel{\text{def}}{=} \text{Cov}(Z,Z) \equiv E[Z^2] - E[Z]^2$

$= a\,\text{Cov}(Z,X) + b\,\text{Cov}(Z,Y)$ from a)

$= a \cdot (a\,\text{Cov}(X,X) + b\,\text{Cov}(X,Y))$ see * below

$+ b \cdot (b\,\text{Cov}(Y,Y) + a\,\text{Cov}(X,Y))$

~~by a) with appropriate variable subs~~ (I think this would work tho)

$\stackrel{\text{def}}{=} a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\,\text{Cov}(X,Y)$ as required ✓

for *, $\text{Cov}(aX+bY+c, X) = E(X \cdot (aX+bY+c)) - E[X]\,E[aX+bY+c]$

$= E[X] \cdot E[bY+c] - E[X]\,E[bY+c] +$ by linearity

and the same for the second term. $aE[X^2] - aE[X]^2 = a\,\text{Var}(X)$

13. $X \sim \text{Bern}(p)$

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

a) $\mathbb{E}[\bar{X}_n] \overset{\text{def}}{=} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right]$

$= \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$

$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i]$ } by linearity

$= \frac{1}{n} \cdot \sum_{i=1}^{n} p$ as each X_i is iid taking 1 w.p. p and 0 otherwise

$= p$

$\text{Var}(\bar{X}_n) \overset{\text{def}}{=} \mathbb{E}[\bar{X}_n^2] - p^2$

$= \mathbb{E}\left[\frac{1}{n^2} \cdot (X_n\right.$

$\text{Var}(\bar{X}_n) \overset{\text{def}}{=} \text{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^{n} X_i\right)$

and as each X_i is iid, $= \frac{1}{n^2} \cdot n \cdot \sigma^2$ where

$\sigma^2 := \text{Var}(X_i) = \text{Var}(X) \overset{\text{def}}{=} \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2$

so $\text{Var}(\bar{X}_n) = \frac{1}{n} \cdot p \cdot (1-p)$

b) The CLT states that $\lim_{n \to \infty} Z_n \sim N(0,1)$

where $Z_n := \frac{\bar{X}_n - \mathbb{E}(\bar{X}_n)}{\sqrt{\text{Var}(\bar{X}_n)}}$ is the standardised sample mean of size n.

i.e. it is normally distributed with mean 0 and variance 1.

c) Let L be the number of failed lightbulbs.

$P(15 \leq L \leq 22) = P(L \leq 22) - P(L \leq 15)$

$P(L \leq k) = P\left(\bar{X}_n \leq \frac{k}{100}\right) = P\left(Z \leq \frac{k/100 - 0.2}{\sqrt{0.2 \cdot 0.8}}\right)$ where $Z \sim N(0,1)$

$= 0.934 - \quad = 0.835$

~~Let $\hat{p}$ be the observed proportion of failed tiptoe buttons.~~

c) ~~$P(0.15 \le \hat{p} \le 0.22) =$~~ we want $P(0.15 \le \bar{X}_{100} \le 0.22)$

where $\mu = 0.2$, $\sigma^2 = \frac{1}{100} \times 0.2 \times 0.8 = 0.0016$, $\sigma = 0.04$

$$= P\left(\frac{0.15 - 0.2}{0.04} \le Z \le \frac{0.22 - 0.2}{0.04}\right)$$

$$= P(Z \le 0.5) - P(Z \le -1.25) = 0.691 - 0.106$$

$$= 0.59$$

15 a) $H_0: \bar{p}_x - \bar{p}_y = 0 \quad H_1: \bar{p}_x - \bar{p}_y > 0$

where $\bar{p}_{\{x,y\}}$ is the proportion of patients recovering who {do, do not} receive the drug. (population)

I'd avoid the bars — usually used for sample means!

~~In our sample,~~ The z statistic

$$z = \frac{(\hat{p}_x - \hat{p}_y) - 0}{\sqrt{\mathrm{Var}(\hat{p}_x - \hat{p}_y)}}$$

where $\hat{p}_x = 0.75$, $\hat{p}_y = 0.65$

and $\mathrm{Var}(\hat{p}_x - \hat{p}_y)$ under H_0 = ~~error~~

$$\frac{\hat{\pi}(1-\hat{\pi})}{100} \cdot 2 \quad \text{with } \hat{\pi} = 0.7$$

By CLT and with a large iid sample, $Z \sim N(0,1)$ given H_0

For a 1-tailed test, the critical values at the 1% and 10% significance levels are 1.282 and 2.326 (satisfying $P(Z \ge C_\alpha) = \alpha$.)

As observed, $z = \frac{0.1}{\sqrt{0.0042}} = 1.543$.

$1.543 < 2.326$, so: insufficient evidence to reject H_0 at 1%

but $1.543 > 1.282$, so we can reject H_0 at 10%.

b) The p-value tells us the smallest significance level α at which we can reject H_0. In this case the p-value, i.e. $P(z \geq 1.543) = 6.14\%$. So we could reject H_0 at 10% but not 1%.

c) A Type I error is when you conclude falsely there's an effect, w.p. $P(\text{reject } H_0 \mid H_0 \text{ is true}) = \alpha$, the significance level. Type II is when you conclude falsely there's no effect, w.p. $P(\text{fail to reject} \mid \text{false}) = \beta$,

where $1-\beta$ is the power of a test.

We'd probably prefer to avoid a Type I error, since prescribing an ineffective drug could be very harmful, but it depends a lot on side-effects, disease severity, alternatives, etc.

d) With $n = 300$, $z = \frac{0.1}{\sqrt{0.0014}} = 2.67$ and now the result is significant at 1%. This isn't especially surprising, more trial participants increases power, as the variance in the different recovery rates is smaller with our larger sample, so we can be more confident the observed difference wasn't too noise.

Great