

ECON 3412 HW 1

David Chen, dc3451

September 29, 2020

Problem 1

a

The sample means are

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum X_i = \frac{1}{6}(1016.6 + 1089.9 + 827.7 + 1021.3 + 1493.3 + 720.1) = 1028.15 \\ \bar{Y} &= \frac{1}{n} \sum Y_i = \frac{1}{6}(320.0 + 448.4 + 575.4 + 219.5 + 561.3 + 195.7) = 386.72\end{aligned}$$

b

I assume by s_X, s_Y that the sample standard deviation is wanted. (All the subsequent parts, when relevant, are computed as the sample covariance, sample correlation, etc.)

$$\begin{aligned}s_X &= \sqrt{\frac{1}{n-1} \sum (X_i - \bar{X})^2} = \sqrt{\frac{1}{5} \sum (X_i - 1028.15)^2} = 266.62 \\ s_Y &= \sqrt{\frac{1}{n-1} \sum (Y_i - \bar{Y})^2} = \sqrt{\frac{1}{5} \sum (Y_i - 386.72)^2} = 166.60\end{aligned}$$

c

The covariance is given by

$$s_{XY} = \frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{5} \sum (X_i - 1028.15)(Y_i - 386.72) = 21590.7$$

d

The sample correlation is given by

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{21590.7}{266.62 \cdot 166.60} = 0.486$$

e

We have that

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} = \frac{21590.7}{266.62^2} = 0.3037$$

f

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 386.72 - (0.3037)1028.15 = 74.44$$

g

We have that each $\hat{Y}_i = 74.44 + 0.3037X_i$, so

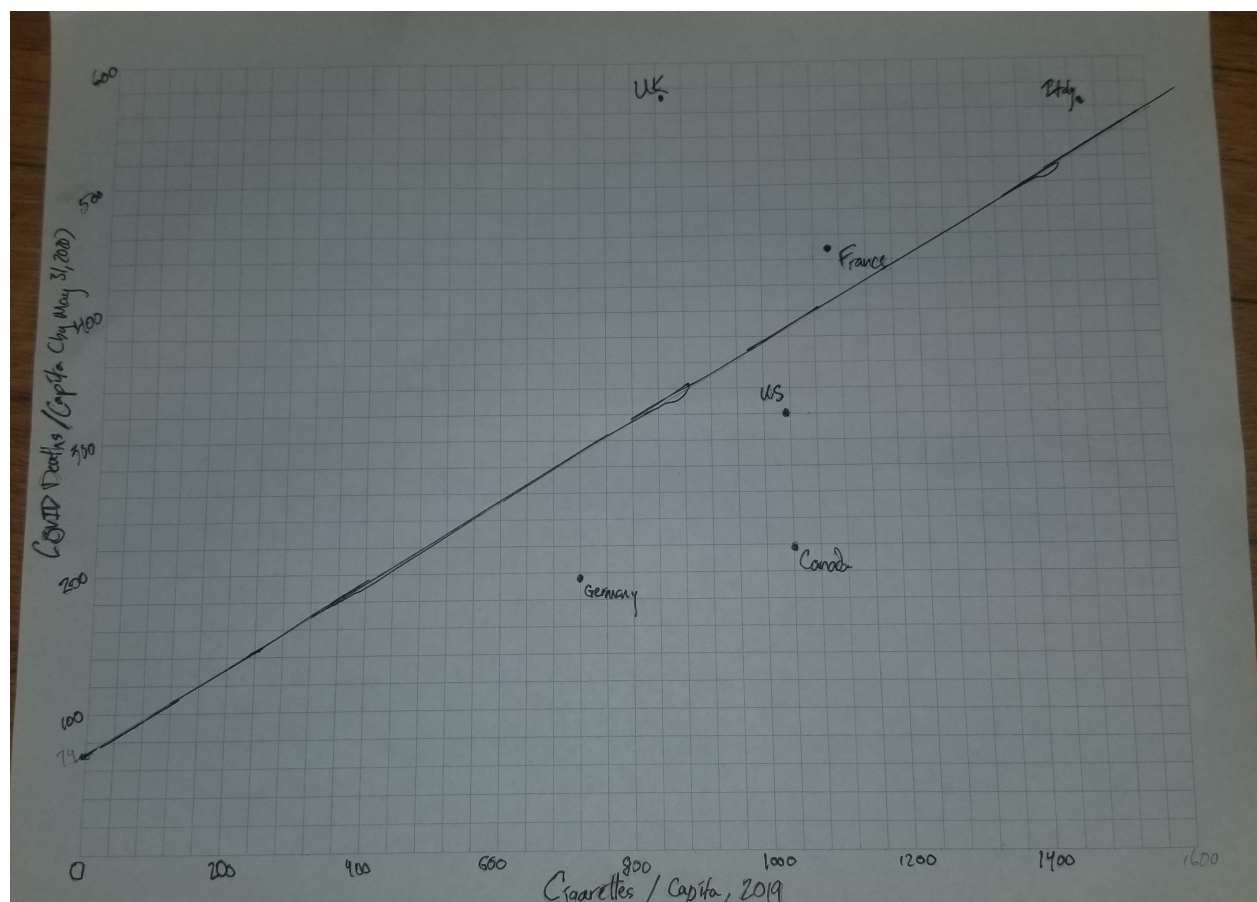
\hat{Y}_1	383.18
\hat{Y}_2	405.44
\hat{Y}_3	325.81
\hat{Y}_4	384.61
\hat{Y}_5	527.96
\hat{Y}_6	293.13

h

We have that each $\hat{u}_i = Y_i - \hat{Y}_i$, so

\hat{u}_1	63.18
\hat{u}_2	-42.96
\hat{u}_3	-249.59
\hat{u}_4	165.11
\hat{u}_5	-33.34
\hat{u}_6	97.43

Problem 2



Problem 3

a

Under the assumption that the population in a opposite-sex relationship has no significant difference between the wider male / female populations, the population mean is $\frac{1}{2}(\mu_M + \mu_W)$, as each relationship has a man and a woman in this case, so we want $E\left[\frac{1}{2}(M + W)\right] = \frac{1}{2}(\mu_M + \mu_W)$ where M, W are the height distributions for men and women respectively.

b

In particular, simple random sampling yields that $E[M_1] = E[M_2] = \mu_M$ and $E[W_1] = E[W_2] = \mu_W$. Then, since expectation is linear, we have

$$\begin{aligned} E[\overline{H}] &= E \left[\frac{1}{4}(M_1 + M_2 + W_1 + W_2) \right] \\ &= \frac{1}{4}(E[M_1] + E[M_2] + E[W_1] + E[W_2]) \\ &= \frac{1}{4}(2\mu_M + 2\mu_W) = \frac{\mu_M + \mu_W}{2} \end{aligned}$$

which was what we wanted.

c

I did this problem set backwards, so I show that $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$ in problem 5. In general, we can use this to see that

$$\text{Var} \left(\sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i,j=1, j \neq i}^n \text{Cov}(X_i, X_j)$$

In particular, we have shown it for $n = 2$, but if it holds for $n = k$, we have that

$$\begin{aligned} \text{Var} \left(\sum_{i=1}^{k+1} X_i \right) &= \text{Var} \left(\sum_{i=1}^k X_i \right) + \text{Var}(X_{k+1}) + 2 \text{Cov} \left(\sum_{i=1}^k X_i, X_{k+1} \right) \\ &= \sum_{i=1}^{k+1} \text{Var}(X_i) + \sum_{i,j=1, i \neq j}^{k+1} \text{Cov}(X_i, X_j) \end{aligned}$$

In this problem, we have that the only nonzero covariances are $\text{Cov}(M_1, W_1) = \text{Cov}(M_2, W_2) = 0.5$, such that we have

$$\begin{aligned} \text{Var}(\overline{H}) &= \text{Var} \left(\frac{1}{4}(M_1 + M_2 + W_1 + W_2) \right) \\ &= \frac{1}{16} \text{Var}(M_1 + M_2 + W_1 + W_2) \\ &= \frac{1}{16} (\text{Var}(M_1) + \text{Var}(M_2) + \text{Var}(W_1) + \text{Var}(W_2) + 2 \text{Cov}(M_1, W_1) + 2 \text{Cov}(M_2, W_2)) \\ &= \frac{1}{16} (\sigma_M^2 + \sigma_W^2 + 2(0.5) + 2(0.5)) \\ &= \frac{1}{16} (\sigma_M^2 + \sigma_W^2 + 2) \end{aligned}$$

d

Since all covariances are 0, we have that analagously to the last part,

$$\begin{aligned}\text{Var } \tilde{H} &= \text{Var}(\gamma_M M_1 + \gamma_M M_2 + \gamma_W W_1 + \gamma_W W_2) \\ &= \gamma_M^2 \text{Var}(M_1) + \gamma_M^2 \text{Var}(M_2) + \gamma_W^2 \text{Var}(W_1) + \gamma_W^2 \text{Var}(W_2) \\ &= 2\gamma_M^2 \sigma_M^2 + 2\gamma_W^2 \sigma_W^2\end{aligned}$$

Its easy to see that this function is twice differentiable with positive-definite Hessian (as we have that $\frac{\partial^2 f}{\partial \gamma_M^2} = 4\sigma_M^2 > 0$, and $\frac{\partial^2 f}{\partial \gamma_M \partial \gamma_W} = 0$, so all pivots of the Hessian are positive), so it is convex.

Since we want the minimal variance under the condition $2\gamma_M + 2\gamma_W = 1$, we have that we need

$$\mathcal{L}(\gamma_M, \gamma_W, \lambda) = 2\gamma_M^2 \sigma_M^2 + 2\gamma_W^2 \sigma_W^2 - \lambda(2\gamma_M + 2\gamma_W - 1)$$

has vanishing partials. Then,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \gamma_M} &= 4\gamma_M \sigma_M^2 - 2\lambda = 0 \\ \frac{\partial \mathcal{L}}{\partial \gamma_W} &= 4\gamma_W \sigma_W^2 - 2\lambda = 0\end{aligned}$$

This gives that

$$\gamma_M = \frac{\sigma_W^2}{\sigma_M^2} \gamma_W$$

Considering the last partial, we have that

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \lambda} &= -2\gamma_M - 2\gamma_W + 1 = 0 \\ &= -2\left(\frac{\sigma_W^2}{\sigma_M^2} \gamma_W + \gamma_W\right) + 1 \\ &= -2\frac{\sigma_W^2 + \sigma_M^2}{\sigma_M^2} \gamma_W + 1\end{aligned}$$

Which gives that

$$\begin{aligned}\gamma_W &= \frac{\sigma_M^2}{2(\sigma_M^2 + \sigma_W^2)} \\ \gamma_M &= \frac{\sigma_W^2}{2(\sigma_M^2 + \sigma_W^2)}\end{aligned}$$

are the coefficients such that $\tilde{H} = \gamma_M M_1 + \gamma_M M_2 + \gamma_W W_1 + \gamma_W W_2$ has a minimal sample variance (since we know that the function is convex, we know this is a minimum).

Problem 4

a

We have the hypotheses:

$$\begin{aligned} H_0 : \mu_B - \mu_S &= 0 \\ H_1 : \mu_B - \mu_S &\neq 0 \end{aligned}$$

where μ_B, μ_S are the population averages of the height of males and females, respectively, at this age, and H_1 is the alternative hypothesis.

b

An unbiased estimate of the difference between μ_B and μ_S by $\overline{Y_{Boys}} - \overline{Y_{Girls}}$.

In this particular sample, $\overline{Y_{Boys}} - \overline{Y_{Girls}} = 57.8 - 58.4 = -0.6$.

c

The variance of the estimator, if we know the population variances σ_{Boys}^2 and σ_{Girls}^2 , will be, as each sampling is assumed to be i.i.d, is

$$\begin{aligned} \text{Var}(\overline{Y_{Boys}} - \overline{Y_{Girls}}) &= \text{Var}(\overline{Y_{Boys}}) + \text{Var}(\overline{Y_{Girls}}) \\ &= \text{Var}\left(\frac{1}{n_{Boys}} \sum_{i=1}^{n_{Boys}} Y_{Boys,i}\right) + \text{Var}\left(\frac{1}{n_{Girls}} \sum_{i=1}^{n_{Girls}} Y_{Girls,i}\right) \\ &= \frac{1}{n_{Boys}^2} \text{Var}\left(\sum_{i=1}^{n_{Boys}} Y_{Boys,i}\right) + \frac{1}{n_{Girls}^2} \text{Var}\left(\sum_{i=1}^{n_{Girls}} Y_{Girls,i}\right) \\ &= \frac{\sigma_{Boys}^2}{n_{Boys}} + \frac{\sigma_{Girls}^2}{n_{Girls}} \end{aligned}$$

Then, the variance of that estimator can be estimated; in particular, we know that the standard deviation of the sampling distribution is s_{Boys}^2/n_{Boys} and s_{Girls}^2/n_{Girls} for $\overline{Y_{Boys}}$ and $\overline{Y_{Girls}}$ respectively. Further, since we know that $\overline{Y_{Boys}}$ and $\overline{Y_{Girls}}$ are constructed from different

randomly selected sample, they are independent, and thus we can estimate the variance of the earlier estimator to be $\frac{s_{Boys}^2}{n_{Boys}} + \frac{s_{Girls}^2}{n_{Girls}}$, as we know the sampling variance to be a consistent estimator for the population variance.

This comes out to 0.765 in this case.

d

We have that $50 \approx \infty$, so by the law of large numbers we have that $\overline{Y_{Boys}}$ and $\overline{Y_{Girls}}$ are approximately normally distributed, so we have that the t -statistic can be given by

$$t = \frac{\overline{Y_{Boys}} - \overline{Y_{Girls}}}{\sqrt{\frac{s_{Boys}^2}{n_{Boys}} + \frac{s_{Girls}^2}{n_{Girls}}}}$$

In particular, since we know that $\overline{Y_{Boys}} - \overline{Y_{Girls}}$ is approximately distributed with mean $\mu_{Boys} - \mu_{Girls}$ and variance $\sigma_{Boys}^2/n_{Boys} + \sigma_{Girls}^2/n_{Girls}$ where σ_X^2 are the population variances (as it is the difference of two approximately normal distributions), $\frac{\overline{Y_{Boys}} - \overline{Y_{Girls}} - (\mu_{Boys} - \mu_{Girls})}{\sqrt{\sigma_{Boys}^2/n_{Boys} + \sigma_{Girls}^2/n_{Girls}}}$ is then the standard normal distribution.

Assuming the null hypothesis, we would then have by the law of large numbers that t , for large n_{Boys}, n_{Girls} , is standard normally distributed itself, as $\mu_{Boys} - \mu_{Girls} \rightarrow 0$, and $\frac{s_{Boys}^2}{n_{Boys}} + \frac{s_{Girls}^2}{n_{Girls}} \rightarrow \sigma_{Boys}^2/n_{Boys} + \sigma_{Girls}^2/n_{Girls}$. (This isn't rigorous, but the result is true, and the justification is at least *morally* true.)

e

In this case, we have that $t = -0.784$. The critical value is, since we know that t is approximately standard normal, 2.575, or $\Phi^{-1}(0.995)$, as we need that the mass under the tails on either end of the distribution total to 0.01. Thus, the difference is not significant.

Intuitively, if we had used a one-sided test, we would have allowed all the mass to come from one side instead of both, and as the critical value is the value such that the mass under the tails beyond the critical value is 0.01, we would've shifted the critical value closer to the mean of the standard normal distribution, i.e. 0, and thus have lowered it. In particular, for a onesided test we would have used $\Phi^{-1}(0.99) = 2.32$.

f

A 95% confidence interval can be given by

$$(\overline{Y_{Boys}} - \overline{Y_{Girls}}) \pm 1.96 \sqrt{\frac{s_{Boys}^2}{n_{Boys}} + \frac{s_{Girls}^2}{n_{Girls}}} = -0.6 \pm 1.96(0.765) = (-2.10, 0.90)$$

Problem 5

a

We have that expectation is linear, such that

$$E[\bar{Y}] = E\left[\frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)\right] = \frac{1}{4}(E[Y_1] + E[Y_2] + E[Y_3] + E[Y_4]) = \frac{1}{4}(4\mu) = \mu$$

The variance satisfies for a, b constants, X, Y some random variables,

$$\begin{aligned}\text{Var}(a + bX) &= E[(a + bX)^2] - E[a + bX]^2 \\ &= a^2 + 2abE[X] + b^2E[X^2] - (a + bE[X])^2 \\ &= a^2 + 2abE[X] + b^2E[X^2] - a^2 - 2abE[X] - b^2E[X]^2 \\ &= b^2(E[X^2] - E[X]^2) = b^2 \text{Var}(X) \\ \text{Var}(X + Y) &= E[(X + Y)^2] - E[X + Y]^2 \\ &= E[X^2] + E[Y^2] + 2E[XY] - E[X]^2 - E[Y]^2 - 2E[X]E[Y] \\ &= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 + 2(E[XY] - E[X]E[Y]) \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)\end{aligned}$$

If X, Y are independent, such as in i.i.d variables, then $\text{Cov}(X, Y) = 0$, and

$$= \text{Var}(X) + \text{Var}(Y)$$

Thus,

$$\begin{aligned}\text{Var}(\bar{Y}) &= \text{Var}\left(\frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)\right) \\ &= \frac{1}{16}(\text{Var}(Y_1 + Y_2 + Y_3 + Y_4)) \\ &= \frac{1}{16}(4\sigma^2) \\ &= \frac{1}{4}\sigma^2\end{aligned}$$

b

$$\begin{aligned}E[\tilde{Y}] &= E\left[\frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4\right] \\ &= \frac{1}{8}E[Y_1] + \frac{1}{8}E[Y_2] + \frac{1}{4}E[Y_3] + \frac{1}{2}E[Y_4] \\ &= \frac{1}{8}\mu + \frac{1}{8}\mu + \frac{1}{4}\mu + \frac{1}{2}\mu = \mu\end{aligned}$$

so \tilde{Y} is unbiased.

The variance is

$$\begin{aligned}\text{Var}(\tilde{Y}) &= \text{Var}\left(\frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4\right) \\ &= \frac{1}{64}\text{Var}(Y_1) + \frac{1}{64}\text{Var}(Y_2) + \frac{1}{16}\text{Var}(Y_3) + \frac{1}{4}\text{Var}(Y_4) \\ &= \frac{11}{32}\sigma^2\end{aligned}$$

c

We have that both \bar{Y} and \tilde{Y} are unbiased estimators, but \tilde{Y} has a higher variance (as $\frac{1}{4} < \frac{11}{32}$). In that sense, since we want to be less susceptible to unlucky sampling (in particular, on average, \tilde{Y} is further from the true mean than \bar{Y}), we want to use \bar{Y} over \tilde{Y} .

d

A feature of the normal distribution is that two normally distributed variables X, Y (which are independent) have that $Z = X + Y$ is also normally distributed with $E[Z] = E[X + Y]$ and $\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y)$. In particular, we also have that each of aY_i is normal with $E[aY_i] = a\mu$ and variance $\text{Var}(aY_i) = a^2\sigma^2$.

Then, this tells us that \bar{Y} and \tilde{Y} , as the sums of normally distributed independent variables, are both normally distributed themselves. The variances and expectations remain unchanged from earlier, with $E[\bar{Y}] = E[\tilde{Y}] = \mu = 5$ and $E[\bar{Y}] = \frac{1}{4}\sigma^2 = \frac{3}{4}$ and $E[\tilde{Y}] = \frac{11}{32}\sigma^2 = \frac{33}{32}$.

Problem 6

a

$$E[GPA \mid SAT = 1600] = 0.9 + 0.001(1600) = 0.9 + 1.6 = 2.5$$

b

$$E[GPA \mid SAT = 2200] = 0.9 + 0.001(2200) = 0.9 + 2.2 = 3.1$$

c

The law of total expectation and the linearity of expectation gives

$$E[GPA] = E[E[GPA \mid SAT]] = E[0.9 + 0.001SAT] = 0.9 + 0.001E[SAT] = 0.9 + 0.001(2000) = 2.9$$

Problem 7

a

The law of total expectation gives us that

$$E[u] = E[E[u \mid X]] = E[0] = 0, E[u^2] = E[E[u^2 \mid X]] = E[\sigma^2] = \sigma^2$$

This gives that the mean of u is $E[u] = 0$ and the variance is $E[(u - E[u])^2] = E[u^2] - E[u]^2 = \sigma^2 - 0 = \sigma^2$.

b

The covariance vanishes if we have that $E[u \mid X] = E[u]$:

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

Via linearity of expectation,

$$= E[uX] - E[u]E[X]$$

Via law of total expectation,

$$= E[E[uX \mid X]] - E[u]E[X]$$

Since X is a constant when conditioned on itself,

$$= E[XE[u \mid X]] - E[u]E[X]$$

We showed earlier that $E[u] = E[u \mid X] = 0$

$$\begin{aligned} &= E[XE[u]] - E[u]E[X] \\ &= E[X]E[u] - E[u]E[X] \\ &= 0 \end{aligned}$$

In particular, this only relies on $E[u \mid X] = k$ for some constant, not necessarily 0.