Bayesian Reasoning and Estimation Theory

1.1 We are considering the probability of the patient having Covid given they are sick.

$$P(\text{Covid} \mid \text{Sick}) = \frac{P(\text{Sick} \mid \text{Covid})\, P(\text{Covid})}{P(\text{Sick})}$$

$P(S \mid C) \rightarrow$ If you have covid you have the symptoms given
$= 1$

$P(C) \rightarrow$ The question says covid has a prevalence of 1%.
$= 0.01$

$P(S) \rightarrow$ If we assume we only get our set of symptoms from covid or the common cold, then we can have a 1% chance of covid, or a 20% chance of the cold. I'm not sure if the two are mutually exclusive.

$$P(S) = P(CC) + P(C) - \underbrace{P(CC \,\&\, C)}$$

If we can only have the cold <u>or</u> covid

$= 0.21$ or 0.208
(exclusive)

Then we plug in to get

$P(S \mid C) = 0.0476$ or 0.0481

Which we'll just round to $0.048 = 4.8\%$

chance of Covid, if we are showing the symptoms shared with the common cold.

1.2 $$P(C|+) = \frac{P(+|C)\,P(C)}{P(+)}$$

We have 2 ways to get a $\oplus$ test

$$P(C|+) = \frac{P(+|C)\,P(C)}{P(+|C)\,P(C) + P(+|\sim C)\,P(\sim C)}$$

$P(\sim C) \rightarrow 1 - P(C)$

For our test, we have

$P(+|C) = 1$ (no false negatives)

$P(+|\sim C) = 0.05$ (5% false positives)

If we don't know about the symptoms they are showing then $P(C) = 0.01$, but it sounds like this is sequential, so we already know about the symptoms they're showing, and $P(C) = 4.8\%$ from 1.1.

$P(C) = 0.048$

Plug everything in:

$$\boxed{P(C|+) = 0.502}$$

(0.17 if we can't use the symptoms evidence, $P(C) = 0.01$)

1.3 Apply some brute force.

Nothing has really changed in the formula, but now $P(C)$ needs to be updated.

We now think that $P(C) = 0.50$

$$P(C \mid ++) = 0.95$$

Update again. $P(C) = 0.95$

$$P(C \mid +++) = 0.997$$

Tests + Symptoms	P(has covid)
0 (symptoms only)	4.8 %
1	50 %
2	95 %
3 (	99.7 %
Tests, no symptom knowledge	
0	1 %
1	17 %
2	80 %
3	99 %

1.4 Let the prevalence estimate be

$$\hat{\theta} = \frac{\oplus \text{ tests}}{\# \text{ of tests}} = P(C)$$

Account for imperfect testing

$$\hat{\theta} = P(+|C)\,P(C) + P(+|\sim C)\,P(\sim C)$$

$$P(+|C) = 1, \quad P(+|\sim C) = 0.05$$

$$\hat{\theta} = P(+|C)\,\theta + P(+|\sim C)(1-\theta)$$

$$\theta = \frac{\hat{\theta} - P(+|\sim C)}{P(+|C) - P(+|\sim C)}$$

For $\hat{\theta} = \frac{100}{1000} = 0.1$ $\theta = 0.0526 = 5.26\%$ prevalence of covid

$\theta = \frac{10}{1000} = 0.01$ $\theta < 0$, which I would interpret as there being no covid prevalence.

We basically got lucky with the tests. I'd expect $P(+|\sim C) \times 1000 = 50$ false positives if there was no covid. I have fewer $\oplus$s than that.

For the uncertainty, we can use a binomial distribution. The variance is $\sigma^2 = npq = np(1-p)$

The probability of testing $\oplus$ is

$$p = P(+|C)\underbrace{P(C)}_{\theta} + P(+|\sim C)\underbrace{P(\sim C)}_{1-\theta} = 0.10$$

$$\Delta\theta^2 = \frac{\sigma^2}{N^2} = 0.09 \qquad \frac{\Delta\theta}{N} = 3\times10^{-4}$$

$\boxed{\theta_{100} = 0.0526 \pm 0.0003}$

Frac of people with covid

Estimate Accuracy.

2.1 I'm going to assume the distribution is Gaussian shaped.

$$P = Norm(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The accuracy is $\Delta\mu_{est}^2 = \frac{\sigma^2}{N} = \frac{Var}{N}$

Our variance is $30\ cm^2 \longrightarrow \Delta\mu_{est} = 0.5\ cm$

2.2 I believe the sociologist should be comfortable concluding that, on average, pinchers are taller than peelers. The uncertainties are not large enough for the means to overlap.

2.3 Because the Gaussians overlap very closely, I don't think they can make any conclusions that dietary habits affect height. We are in the Rayleigh limit, where our Gassian's are combining. If we look at the whole population data, it's just one large Gaussian. We can't distinguish between Peelers and Pinchers.

In order to improve our data so we can make inferences, I would recall the superresolution

2.3 (cont.) problem. There, the key ingredient was finding the probability of collecting a photon from each source. Analagously, if we knew the prevalence of Peelers and Pinchers, we might be able to differentiate between the two.

$$P(\text{pincher}|\text{height}) = \frac{P(\text{pincher})\,P(\text{height}|\text{pincher})}{P(\text{height})}$$

After rereading, I see it's a 50/50 split, so $P(\text{pinch}) = P(\text{peel})$

Then: $$P(\text{pinch}|h) = \frac{\frac{1}{2}\mathcal{N}(h;\mu_{pinch})}{\frac{1}{2}\mathcal{N}(h;\mu_{pinch}) + \frac{1}{2}\mathcal{N}(h;\mu_{peel})}$$

Then if we're not close to the mean of either, we can tell who is who with better certainty. If we are close to the mean, it's very difficult to make an accurate prediction.

Other factors to consider might be dietary habits besides bananas. Perhaps most of the pincher faction is made up of certain families that have passed down the pincher tradition for many generations. Then the difference may be genetic.

Finally, I'd like to know their prior. ~~Before~~ Before sampling the populations, did they expect how someone eats a banana to affect their height? If that prior is small, then we can probably wave away a small differente in mean heights.