Combining statistics by gender for equal opportunity hiring

I was once asked by the HR department how to address the following statistical problem:

> *We have the answers from a candidate of a personality test, producing a score. We have some statistical information tables about the score distribution for each gender, but we are not allowed to ask it. How should we combine information from both genders to interpret the candidate score?*

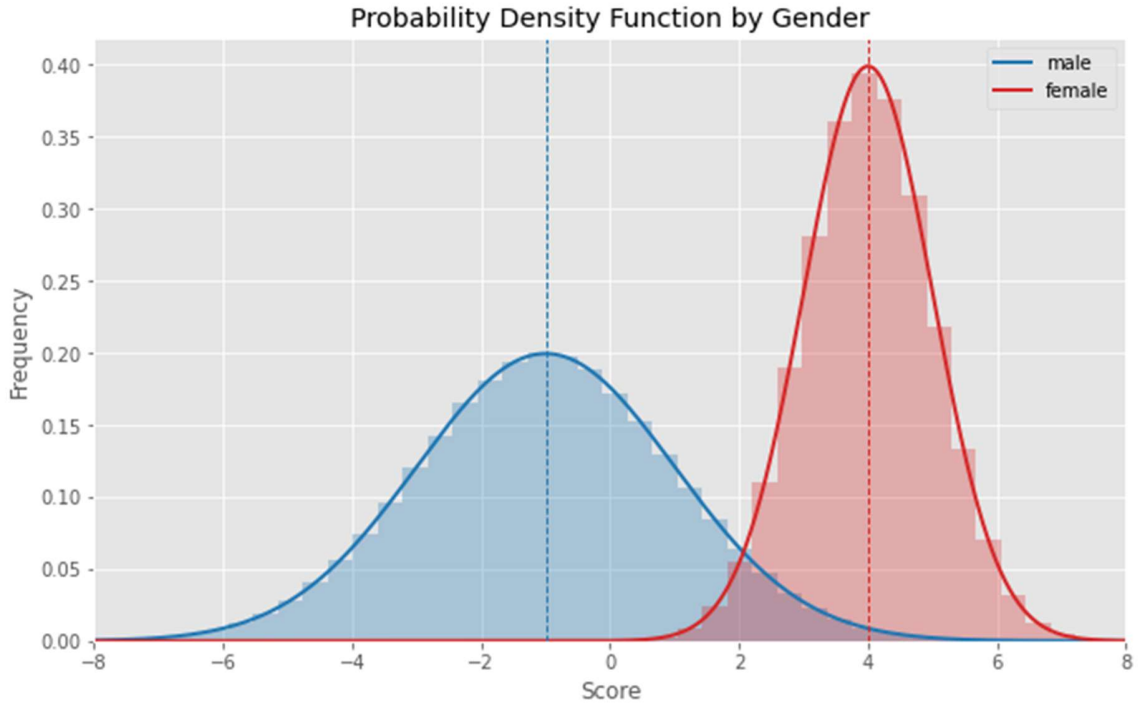
Image by Arek Socha from Pixabay

Short answer: average the percentile of the score for both genders.
Yes, that simple. Here's how 👇.

Consider an exaggerated example of two very different normal distributions.

	mean	std
male	-1	2
female	4	1

Let us draw them both theoretically (*solid lines*) and by sampling (*histograms*).



Common mistake: average of normals¶

Assuming each gender's score is normally distributed with known mean and variance,

$$X \sim N(\mu_m, \sigma_m^2)$$
$$Y \sim N(\mu_f, \sigma_f^2)$$

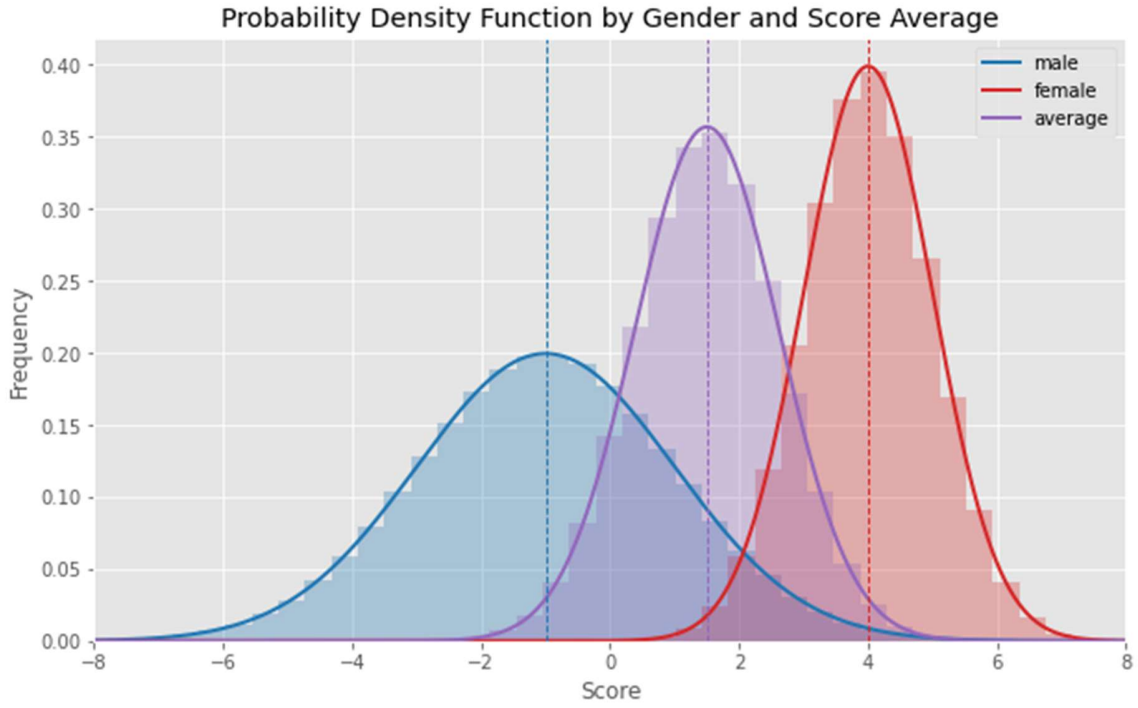
use the formula for average of normals, which is the normal with mean of means and quadratic mean of standard deviations,

$$Z = \frac{X+Y}{2} \sim N(\frac{\mu_m + \mu_f}{2}, \frac{\sigma_m^2 + \sigma_f^2}{4})$$

Assuming a normal distribution for each gender is usually not incorrect. In fact, we are usually given only the sample mean and standard deviation as only descriptors. The mistake is to confuse a mixture of normals with an average of normals.

An average of normals would be valid if, for example, we made **pairs of male and female candidates and averaged their score.** The resulting distribution would also be normal, a nd inexperienced statisticians will mistakenly be compelled towards this desired property.

Let us draw the average of normals, just for comparison.



Correct approach: mixture of normals

Our case is different. We read a score from a candidate, which can be male with probability *p(male)* or female with probability *p(female)=1−p(male)*, ie, gender is a *Bernouilli(p(male)).* This phenomena is called **mixture** of normals, and we can model it via the density of probabilities.

The probability of a candidate having score *x* can be obtained as a weighted sum of probabilities assuming each gender, the weights being the probability of belonging to each genders. This is just an application of the total probability formula:

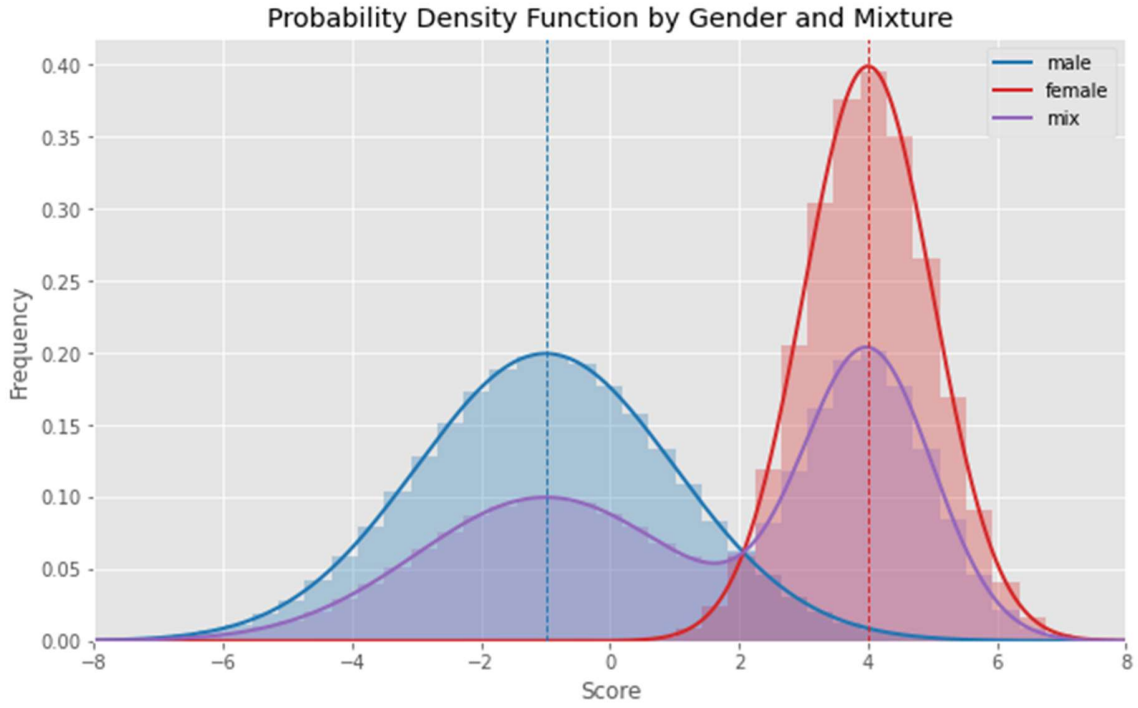
$$p(x) = p(\text{male})p(x|\text{male}) + p(\text{female})p(x|\text{female})$$

Unless prior knowledge is available, we will assume there are the same number of male as female candidates, so *p(male)=p(female)=1/2*.

$$p(x) = \frac{1}{2}p(x|\text{male}) + \frac{1}{2}p(x|\text{female})$$

$$X|\text{male} \sim \mathcal{N}(\mu_m, \sigma_m^2)$$
$$X|\text{female} \sim \mathcal{N}(\mu_f, \sigma_f^2)$$

One way to think about it is that the averaging happens not at the variable level, but on the probability space.



Note that this is not a normal distribution, it is bimodal for sufficiently apart distributions.

From probability densities to percentiles

Finally, a common way to position a score inside its distribution is the percentile, that is, out of 100 people, how does one's score rank: 0 is the minimum, 50 the median and 100 the maximum. The cumulative probability of a variable X at a value x is defined as the probability of X having a value lower or equal than x, and it can be expressed as an integral of the point-wise density function below x

$$F_X(x) = P(X \leq x) = \int^x p(x)dx$$

, and it is well known and available for common distributions, including the normal. If we multiply it by 100 we obtain such ranking

$$\text{Percentile}(x) = 100F(x)$$

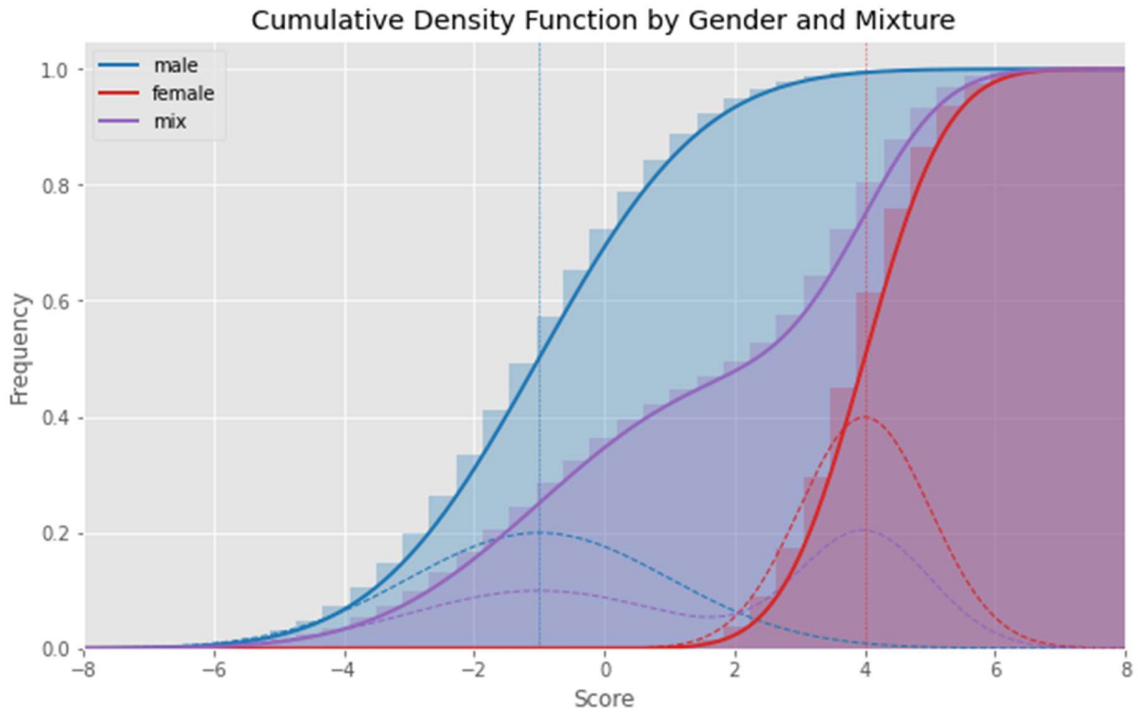
If we now decompose the point-wise density function for the mixture as an average of both male and female probabilities, since the integral of averages is the average of integrals,

$$F_X(x) = P(X \leq x) = \int^x p(x)dx = \int^x \left(\frac{1}{2}p(x|\text{male}) + \frac{1}{2}p(x|\text{female})\right) dx$$
$$= \frac{1}{2}\int^x p(x|\text{male})dx + \frac{1}{2}\int^x p(x|\text{female})dx$$
$$= \frac{1}{2}F_{\text{male}}(x) + \frac{1}{2}F_{\text{female}}(x)$$

So assuming we can compute the cumulative density function (cdf) for each one, the percentile can be computed using just the average of both, for note that

$$\text{Percentile}(x) = \frac{1}{2}\text{Percentile}_{\text{male}}(x) + \frac{1}{2}\text{Percentile}_{\text{female}}(x)$$

Note: Combining percentiles is valid for any distribution, not only normals, including tabulated values.



So, for example, a candidate scoring x=2.5, would rank

- ~5% among female population
- ~95% among male population
- ~50% among the total population, ie, if gender is unknown.

We provide a function in python for computing combined percentiles from two normal distributions with known parameters.

```
import scipy.stats as st

def percentile_mix(x, m1, s1, m2, s2, ratio1=1/2, ratio2=1/2):
  percent1=st.norm(m1,s1).cdf(x)*100
  percent2=st.norm(m2,s2).cdf(x)*100
  percent_mix = (ratio1*percent1+ratio2*percent2)/(ratio1+ratio2)
  return percent_mix
```

If you are using *Microsoft Excel*, you can use

=NORM.DIST(x, mean, standard_dev, TRUE)

for the percentile of score value x in each normal distribution, and then take the average of both.