Statistics : Solutions

David W.H. Swenson

Exercise 1. If you roll this die 25 times, about how many times will you expect to get each value (1, 2, and 3)?

For each possible result, we have to multiple its probability (from the table) by the number of trials (25 in this problem). Since "how many times" requires an integer response, we give our responses as integers.

Number Rolled	Number of Times
1	$0.167 \times 25 \approx 4.2$, so 4 or 5
2	$0.333 \times 25 \approx 8.3$, so 8 or 9
3	$0.500 \times 25 = 12.5$, so 12 or 13

Exercise 2. The height distribution function shown above is generated from the function

$$P(x) = \frac{1}{944.998}\left(x\left(e^{-0.095(x-64)^2} + 1.1\, e^{-0.055(x-69.4)^2}\right)\right) \tag{1}$$

What is the probability that someone is shorter than 60 inches? Taller than 76 inches? Between 64 inches and 69 inches? [Hint : you'll probably want to use a calculator or computer program to do the numeric integration for you.]

For each of these, we just need to take the appropriate integrals. In each of the following integrals, $P(x)$ is given by equation 1.

1. Shorter than 60 inches:

$$\int_0^{60} \mathrm{d}x\, P(x) = 0.015$$

   Remember height can't be negative, so 0 is the shortest height possible.

2. Taller than 76 inches:

$$\int_{76}^{\infty} \mathrm{d}x\, P(x) = 0.010$$

   People can be as tall as possible, all the way up to infinity! (But if you're doubtful, figure out the percentage of people taller than 8 ft, or 96 inches.) But what if your numerical integrator (maybe a TI-83 calculator) doesn't have a way to represent infinity? How do you go all the way out to infinity?

Since the probability distribution function rapidly falls off to zero, you could just do the integration up to some arbitrary large number (like 200) and trust that to be good to as many decimal places as we have. Another method would be to take the value of the integral from zero to infinity (see exercise 8) and subtract off the numerically-calculated integral from 0 to 76.

3. Between 64 and 69 inches (average height of women and average height of men):

$$\int_{64}^{69} \mathrm{d}x\, P(x) = 0.436$$

Just in case anyone is curious, that PDF is not based on anything realistic. I pretty much just made it up out of thin air.

Exercise 3. What would be wrong with using $P(x) = x^2$ as a probability distribution function over the domain of all real numbers? [Hint : what is the probability for a value between 0 and 2?]

Following the hint, we find the probability for a value between 0 and 2:

$$\int_0^2 \mathrm{d}x\, x^2 = \frac{1}{3}x^3\Big]_0^2 = \frac{1}{3}2^3 - 0 = 8/3 \approx 2.67$$

So the probability is 2.67. But since probability can't be higher than one, this clearly doesn't make sense — it suggests more than 100% probability.

Exercise 4. What would be wrong with using $P(x) = x^3$ as a probability distribution function if the domain was restricted to $[-1, 1]$? [Hint : what is the probablity for a value between -1 and 0?]

Again taking the hint, we find that the probability for a value between -1 and 0 is

$$\int_{-1}^0 x^3 = \frac{1}{4}x^4\Big]_{-1}^0 = 0 - \frac{1}{4} = -0.25$$

As with the previous problem, we have a solution that doesn't make sense. A probability can not be negative.

Exercise 5. From equations 3 and 4, show that

$$\sum_{j=1}^{M} p_j = 1 \tag{2}$$

That is, show that the sum of all the probabilities p_j is equal to one.

From equation 4, we have

$$\sum_{j=1}^{M} p_j = \sum_{j=1}^{M} \frac{n_j}{N} = \frac{1}{N}\sum_{j=1}^{M} n_j$$

But then equation 3 gives us

$$\sum_{j=1}^{M} p_j = \frac{1}{N}\sum_{j=1}^{M} n_j = \frac{1}{N}N = 1$$

Exercise 6. Show that equation 6 implies that $\int_I \mathrm{d}x\, P(x) = 1$.

Once again, we just plug in the equation:

$$\int_I \mathrm{d}x\, P(x) = \int_I \mathrm{d}x\, \frac{p(x)}{A} = \frac{1}{A}\int_I \mathrm{d}x\, p(x) = \frac{1}{A}A = 1$$

Exercise 7. What is the normalization constant A for the Gaussian PDF $p(x) = e^{-ax^2}$, defined over all space? Give your answer in terms of a. [Hint : you'll probably need to look up the definite integral.]

As the hint suggests, this all boils down to doing one integral:

$$A = \int_{-\infty}^{\infty} \mathrm{d}x\, e^{-ax^2} = \sqrt{\frac{\pi}{a}}$$

where the value of the definite integral comes from a table of integrals. Remember that this means that the normalized Gaussian PDF is $P(x) = p(x)/A = \sqrt{\frac{a}{\pi}}e^{-ax^2}$.

Exercise 8 (Advanced)**.** Show that the PDF from exercise 2 is normalized. Note that this is more an exercise in integration than in statistics. You can cheat by doing it approximately numerically, but it would good practice to actually try it by doing the integrals. By using the approximation $\int_{-\infty}^{-b} e^{-ax^2} \approx 0$ for $b > 3.5/a$, you can solve the integrals without using a computer. You'll need to remember that $\int_a^b \mathrm{d}x\, f(x) + \int_b^c \mathrm{d}x\, f(x) = \int_a^c \mathrm{d}x\, f(x)$ and $\int_a^b \mathrm{d}x\, (f(x) + g(x)) = \int_a^b \mathrm{d}x\, f(x) + \int_a^b \mathrm{d}x\, g(x)$. Remember that the bounds of normalization are $[0, \infty]$ (height can't be negative).

For simplicity of notation, we'll make the following definitions:

$$\begin{aligned} A &\equiv 944.998 \\ a &\equiv 0.095 \\ \alpha &\equiv 64 \\ \Gamma &\equiv 1.1 \\ b &\equiv 0.055 \\ \beta &\equiv 69.4 \end{aligned}$$

With those notations, we write the probability distribution function:

$$\begin{aligned}
\int_I \mathrm{d}x\, P(x) &= \int_0^\infty \mathrm{d}x\, \frac{1}{944.998}\left(x\left(e^{-0.095(x-64)^2} + 1.1\, e^{-0.055(x-69.4)^2}\right)\right) \\
&= \int_0^\infty \mathrm{d}x\, \frac{1}{A}\left(x\left(e^{-a(x-\alpha)^2} + \Gamma\, e^{-b(x-\beta)^2}\right)\right) \\
&= \frac{1}{A}\int_0^\infty \mathrm{d}x\, \left(x\, e^{-a(x-\alpha)^2} + \Gamma x\, e^{-b(x-\beta)^2}\right) \\
&= \frac{1}{A}\left(\int_0^\infty \mathrm{d}x\, x\, e^{-a(x-\alpha)^2} + \int_0^\infty \mathrm{d}x\, \Gamma x\, e^{-b(x-\beta)^2}\right)
\end{aligned}$$

Now, with each integral we do a variable substitution. For the first integral, we make the substitution $u = x - \alpha$. For the second, we use $v = x - \beta$. Substituting those back into the expression above, we obtain

$$\begin{aligned}
\int_I \mathrm{d}x\, P(x) &= \frac{1}{A}\left(\int_{-\alpha}^\infty \mathrm{d}u\, (u+\alpha) e^{-au^2} + \Gamma \int_{-\beta}^\infty \mathrm{d}v\, (v+\beta) e^{-bv^2}\right) \\
&= \frac{1}{A}\left(\int_{-\alpha}^\infty \mathrm{d}u\, u\, e^{-au^2} + \int_{-\alpha}^\infty \mathrm{d}u\, \alpha\, e^{-au^2}\right. \\
&\qquad \left. + \Gamma \int_{-\beta}^\infty \mathrm{d}v\, v\, e^{-bv^2} + \Gamma \int_{-\beta}^\infty \mathrm{d}v\, \beta\, e^{-bv^2}\right)
\end{aligned}$$

Now we'll refresh you on the tip given in the problem (which is actually kind of a cheat). To as many digits as we expect to have accuracy (about 5) we can safely make the approximations described in the text. Since $\alpha = 64 > 3.5/a \approx 36.84$ and $\beta = 69.4 > 3.5/b \approx 63.64$, we accept the approximations as true (show-offs can use previously discussed techniques of numerically integrating to infinity to figure out just *how* accurate we are.) So now what we're going to do is add zero to the equation four times, each as some variant on $0 \approx \int_{-\infty}^{-b} \mathrm{d}x\, e^{-ax^2}$.

$$\begin{aligned}
\int_I \mathrm{d}x\, P(x) &= \frac{1}{A}\left(\int_{-\infty}^{-\alpha} \mathrm{d}u\, u\, e^{-au^2} + \int_{-\alpha}^\infty \mathrm{d}u\, u\, e^{-au^2} + \int_{-\infty}^{-\alpha} \mathrm{d}u\, \alpha\, e^{-au^2}\right. \\
&\qquad + \int_{-\alpha}^\infty \mathrm{d}u\, \alpha\, e^{-au^2} + \Gamma \int_{-\infty}^{-\beta} \mathrm{d}v\, v\, e^{-bv^2} + \Gamma \int_{-\beta}^\infty \mathrm{d}v\, v\, e^{-bv^2} \\
&\qquad \left. + \Gamma \int_{-\infty}^{-\beta} \mathrm{d}v\, \beta\, e^{-bv^2} + \Gamma \int_{-\beta}^\infty \mathrm{d}v\, \beta\, e^{-bv^2}\right) \\
&= \frac{1}{A}\left(\int_{-\infty}^\infty \mathrm{d}u\, u\, e^{-au^2} + \int_{-\infty}^\infty \alpha\, e^{-au^2}\right. \\
&\qquad \left. + \Gamma \int_{-\infty}^\infty \mathrm{d}v\, v\, e^{-bv^2} + \Gamma \int_{-\infty}^\infty \mathrm{d}v\, \beta\, e^{-bv^2}\right) \\
&= \frac{1}{A}\left(0 + \alpha \int_{-\infty}^\infty \mathrm{d}u\, e^{-au^2} + 0 + \Gamma\beta \int_{-\infty}^\infty \mathrm{d}v\, e^{-bv^2}\right) \\
&= \frac{1}{A}\left(\alpha\sqrt{\frac{\pi}{a}} + \Gamma\beta\sqrt{\frac{\pi}{b}}\right)
\end{aligned}$$

Plugging in the numbers, we have:

$$\begin{aligned}\int_I \mathrm{d}x\, P(x) &= \frac{1}{944.998}\left(64\sqrt{\frac{\pi}{0.095}} + 1.1 \cdot 69.4\sqrt{\frac{\pi}{0.055}}\right) \\ &\approx \frac{1}{944.998} 944.998 = 1\end{aligned}$$

Exercise 9. Imagine that in the loaded-die example from section 1.1 we rolled the die 30 times, and 5 times we got the number 1, 10 times we got the number 2, and 15 times we got the number 3. In the two formulae above, what is N? What is M? n_1? n_2? n_3? Is there any meaning to n_4? What is χ_j for $j = 2$? Do we know what x_i is for $i = 2$?

N	30 (number of trials)
M	3 (number of possible outcomes)
n_1	5 (number of times we got the result 1)
n_2	10 (number of times we got the result 2)
n_3	15 (number of times we got the result 3)
n_4	0 or undefined (number of times we got the result 4, which never happened because it isn't an option)
χ_j for $j = 2$	2 (value for result indexed by number 2, which conveniently is also the value)
x_i for $i = 2$	unknown (value of the second trial — since we only know how many times each event happened, and not the order in which they happened, we don't know this)

Exercise 10. For some generic function $f(x)$, show that given a discrete probability distribution function p_j, we have:

$$\langle f(x) \rangle = \sum_{j=1}^{M} p_j f(\chi_j)$$

We'll essentially follow the same derivation as used in the text for $\langle x \rangle$, except this time we're looking for the average of $f(x)$ instead of just x.

$$\begin{aligned}\langle f(x) \rangle &= \frac{1}{N}\sum_{j=1}^{M} n_j f(\chi_j) \\ &= \sum_{j=1}^{M} \frac{n_j}{N} f(\chi_j) \\ &= \sum_{j=1}^{M} p_j f(\chi_j)\end{aligned}$$

Exercise 11. In summation notation, write out $\langle f(x) \rangle$ and $f(\langle x \rangle)$. In general, does $\langle f(x) \rangle - f(\langle x \rangle) = 0$?

$$\langle f(x)\rangle = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

$$f(\langle x\rangle) = f\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)$$

In general, these two will not be equal. So we can also say that in general $\langle f(x)\rangle - f(\langle x\rangle) \neq 0$.

Exercise 12. What is the expectation value of x for the normalized Gaussian distribution function $P(x) = \sqrt{\frac{a}{\pi}}e^{-ax^2}$, defined over all space? [Hint: A Gaussian is a bell curve (this one centered at 0). So you can do the first part in your head.] Using that, what is $\langle x\rangle^2$? What about the expectation value $\langle x^2\rangle$?

Since a Gaussian is symmetric about its center (that is, it is an even function about the center), the center is also the expectation value of x for it. Another way of looking at this is to note that $P(x)$ is an even function, and so the expectation value $\langle x\rangle = \int_{-\infty}^{\infty} \mathrm{d}x\, x\, P(x)$ is the integral of an odd function (an even function times an odd function is an odd function) over an even interval. So $\langle x\rangle$ is zero.

$$\begin{aligned}
\langle x\rangle &= 0 \\
\langle x\rangle^2 &= 0^2 = 0 \\
\langle x^2\rangle &= \int_{-\infty}^{\infty} \mathrm{d}x\, x^2 P(x) \\
&= \int_{-\infty}^{\infty} \mathrm{d}x\, x^2 \sqrt{\frac{a}{\pi}} e^{-ax^2} \\
&= \sqrt{\frac{a}{\pi}} \int_{-\infty}^{\infty} \mathrm{d}x\, x^2 e^{-ax^2} \\
&= \sqrt{\frac{a}{\pi}} \sqrt{\frac{\pi}{4a^3}} = \frac{1}{2a}
\end{aligned}$$

Exercise 13. What is the variance of a normalized (over all space) Gaussian, $\sqrt{\frac{a}{\pi}}e^{-ax^2}$? If the Gaussian was not normalized, would that change its variance?

We actually did most of this problem in exercise 12. Since we know that $\langle x\rangle = 0$, we have

$$\sigma^2 = \left\langle (x - \langle x\rangle)^2 \right\rangle = \left\langle (x-0)^2 \right\rangle = \langle x^2\rangle = \frac{1}{2a}$$

Looking at our solution for $\langle x^2\rangle$ above, if the Gaussian was not normalized, we wouldn't have had the factor of $\sqrt{a/\pi}$ to simplify our expression. So normalization affects the variance.

Exercise 14. Show that $\sigma^2 = \langle x^2\rangle - \langle x\rangle^2$. [Hint: $\langle f(x) + g(x)\rangle = \langle f(x)\rangle + \langle g(x)\rangle$ for any functions f and g.]

This basically just requires that we use the hint:

$$\begin{aligned}\sigma^2 &= \left\langle (x - \langle x \rangle)^2 \right\rangle \\ &= \left\langle x^2 - 2x \langle x \rangle + \langle x \rangle^2 \right\rangle \\ &= \left\langle x^2 \right\rangle - \langle 2x \langle x \rangle \rangle + \left\langle \langle x \rangle^2 \right\rangle\end{aligned}$$

Now it's important to remember that $\langle x \rangle$ is just a number, so it can be treated like a constant. Since $\langle c \rangle = c$ for a constant, we have

$$\begin{aligned}\sigma^2 &= \left\langle x^2 \right\rangle - 2 \langle x \rangle \langle x \rangle + \langle x \rangle^2 \\ &= \left\langle x^2 \right\rangle - 2 \langle x \rangle^2 + \langle x \rangle^2 \\ &= \left\langle x^2 \right\rangle - \langle x \rangle^2\end{aligned}$$