6.2 Statistical inference

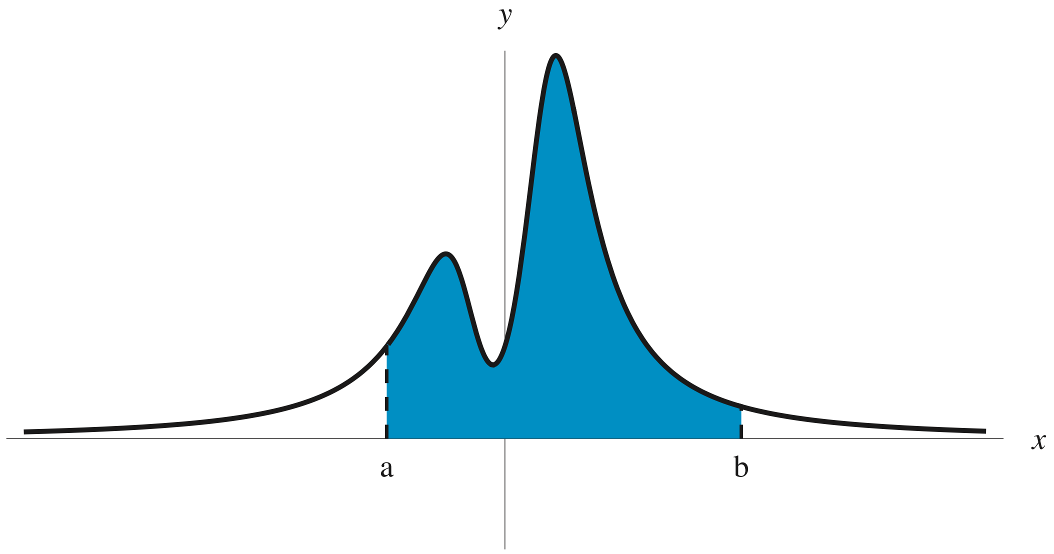
Please note that this section is in "DIY" (Do It Yourself) format: there are various blanks to be filled in, and questions to be answered, *along the way*, rather than being left to a set of exercises at the end. (SOLUTIONS IN RED)

A. Random variables and pdf's

Let X be a random variable, meaning, essentially, a way of assigning a real number to each possible outcome of an experiment. We say that X has *probability density function*, or *pdf*, given by $f(x)$ if

$$P(a < X < b) = \int_a^b f(x)\,dx$$

for any numbers a and b in the domain (set of possible values) of X. (Again, $P(a < x < b)$ denotes the probability that, if a value x is chosen from X at random, that value will lie between the numbers a and b.)



Exercise A1. Fill in the blanks: the mean μ and standard deviation σ of a pdf $f(x)$ can be obtained as follows. Draw a relative frequency __density__ histogram corresponding to a sample of points from X. Compute the __mean__ $\overline{x}$ and the __standard deviation__ s of the histogram data. Repeat for larger and larger samples, using narrower and narrower bin widths. Then the tops of the bars of the histogram will smooth out to give you the graph of your pdf $y =$__$f(x)$__, and your numbers $\overline{x}$ and s will converge to __μ__ and __σ__, respectively.

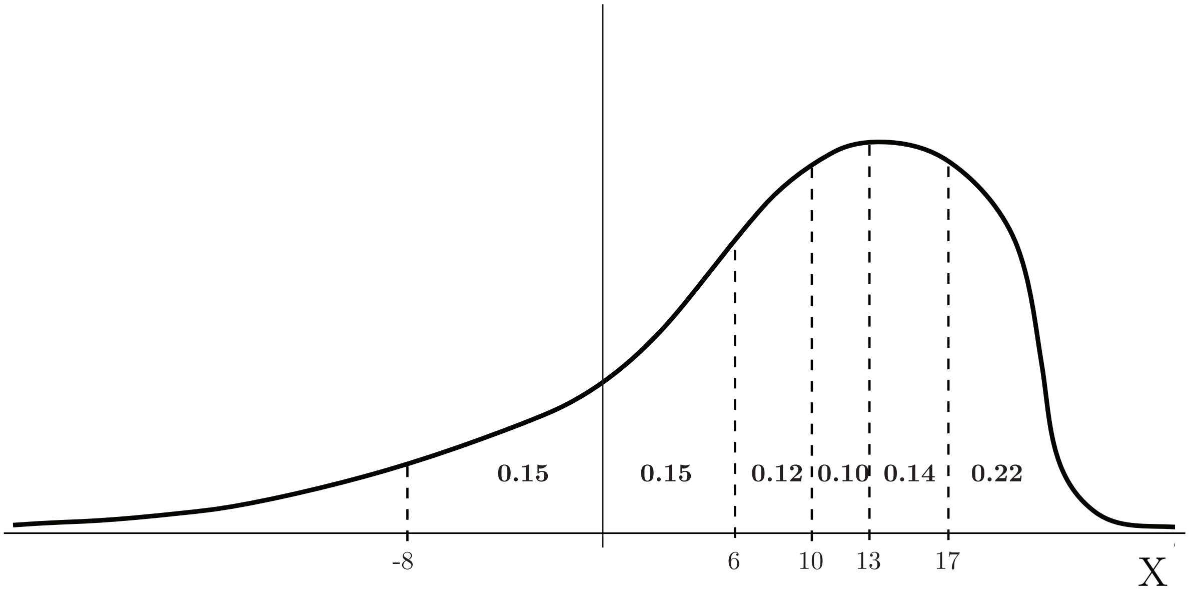
Also recall that, for any pdf $f(x)$, with domain (c,d), we must have

- $f(x) \geq 0$ for all x in (c,d) (since probabilities can't be negative), and

- $\int_c^d f(x)\,dx = 1$ (since the probability that a data point in X lies somewhere in X must equal 100%, or 1).
- $P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b)$ for all a and b in (c,d) (probabilities are the same whether or not you include endpoints, since the area under a single point on the graph of a function is zero).

Remark. Often, the domain (c,d) of a pdf will be taken to be of *infinite* extent, meaning $c = -\infty$ or $d = +\infty$, or both.

Exercise A2. Consider the following probability density function for a random variable X. The regions delineated by dashed lines have areas as shown.



Find:

(a) $P(X < -8)$. 0.12

(b) $P(10 < X < 17)$. 0.24

(c) The number c such that 36% of all data values of X are at least 6 and at most c. $c = 17$

B. Standard normal random variables and pdf's

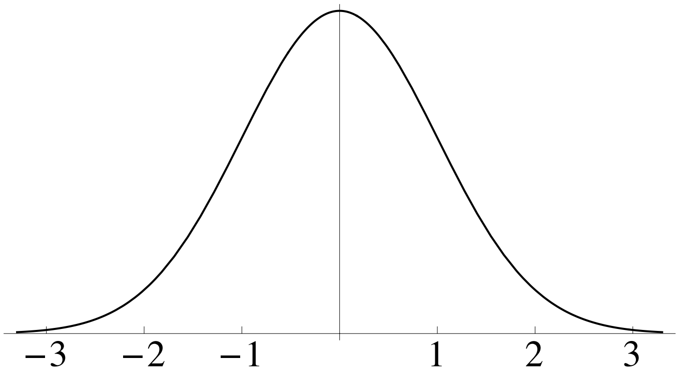
A random variable X is said to have a *standard normal distribution* if the pdf for X is given by

$$f_{0,1}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

(where x can be any real number).

For such an X, we say "X is $N(0{,}1)$." In other words, to say X is $N(0{,}1)$ is to say that, for any real numbers a, b with $a < b$,

$$P(a < X < b) = \int_a^b f_{0,1}(x)\,dx = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-x^2/2}\,dx.$$



FACT: The pdf $f_{0,1}(x)$ has mean $\mu = 0$ and standard deviation $\sigma = 1$. (That's why we call it $f_{0,1}(x)$.)

Exercise B1. Fill in the blanks:

(a) It can be computed that $\int_{-1}^{1} f_{0,1}(x)\,dx \approx 0.683$. In other words: if X is $N(0{,}1)$, then about ___68.3___% of the values of X lie within one ___standard deviation___ of the mean.

(b) It can be computed that $\int_{-2}^{2} f_{0,1}(x)\,dx \approx 0.955$. In other words: if X is $N(0{,}1)$, then about ___95.5___% of the values of X lie within ___two___ standard deviations of the mean.

(c) It can be computed that $\int_{-3}^{3} f_{0,1}(x)\,dx \approx 0.997$. In other words: if X is $N(0{,}1)$, then about ___99.7___% of the values of X lie within three ___standard deviations___ of the mean.

(d) It can be computed that $\int_{-1.96}^{1.96} f_{0,1}(x)\,dx \approx 0.950$. In other words: if X is $N(0{,}1)$, then about ___95___% of the values of X lie within ___1.96___ standard deviations of the mean.

(e) It can be computed that $\int_{-2.33}^{2.33} f_{0,1}(x)\,dx \approx 0.980$. In other words: if X is $N(0{,}1)$, then about ___98___% of the values of X lie within ___2.33___ standard deviations of the mean.

(f) It can be computed that $\int_{-2.576}^{2.576} f_{0,1}(x)\,dx \approx 0.990$. In other words: if X is $N(0{,}1)$, then about ___99___% of the values of X lie within ___2.576___ standard deviations of the mean.

C. Other normal random variables and pdf's

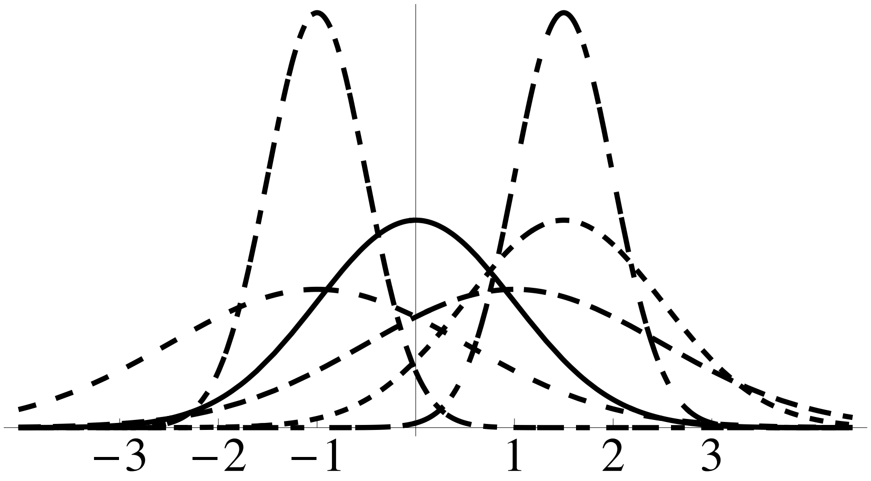
If the pdf for a random variable Z follows the basic shape of the standard normal curve, but has mean μ (instead of 0) and standard deviation σ (instead of 1), we say "Z is $N(\mu,\sigma)$." Let's denote such a pdf by $f_{\mu,\sigma}(x)$. Then: to say Z is $N(\mu,\sigma)$ is to say that, for any real numbers a, b with $a < b$,

$$P(a < Z < b) = \int_a^b f_{\mu,\sigma}(x)\,dx.$$

The precise formula for $f_{\mu,\sigma}(x)$ is

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}.$$

But *we won't need* this formula, because we are going to *translate* $N(\mu,\sigma)$ variables to $N(0,1)$ variables, shortly.



Exercise C1. Recall that the mean of a data set measures the "central tendency," and that the standard deviation measures the "spread" (large standard deviation means large spread, and conversely). Given all this, and also assuming that the solid curve on the graph above is $N(0,1)$, identify, on the above graph, which of the dashed curves is $N(1.5,1)$; which is $N(1,1.5)$; which is $N(1.5,0.5)$; which is $N(-1,1.5)$; and which is $N(-1,0.5)$. Please explain your reasoning briefly, in the space below.

The two tallest curves are the least spread out, so they must have the smallest of the standard deviations, which is 0.5. The leftmost of these taller curves is centered at -1 and therefore has mean -1; the rightmost, similarly, has mean 1.5. Similar arguments apply to the other curves.

D. Translation between $N(0,1)$ variables and $N(\mu,\sigma)$ variables

We have the following NISNID ("Normal Is Standard Normal In Disguise") Fact, which we present without proof (but which is not hard to show, using the above formula for the $N(\mu,\sigma)$ pdf $f_{\mu,\sigma}(x)$):

NISNID Fact. If Z is $N(\mu,\sigma)$, then $\dfrac{Z-\mu}{\sigma}$ is $N(0,1)$.

In stats texts, you will typically find tables of $N(0,1)$ variables, but not other $N(\mu,\sigma)$ variables. Now we know why: we can *compute* probabilities associated with $N(\mu,\sigma)$ random variables if *all we know* are probabilities associated with $N(0,1)$ random variables.

Here's an example showing how.

Example. Suppose Z is $N(8,1.5)$. Find $P(5 < Z < 11)$.

Solution. Since, in this case, $\mu = 8$ and $\sigma = 1.5$, we have

$$\begin{aligned} P(5 < Z < 11) &= P\left(\frac{5-8}{1.5} < \frac{Z-8}{1.5} < \frac{11-8}{1.5}\right) \\ &= P\left(-2 < \frac{Z-8}{1.5} < 2\right) = 0.955. \end{aligned}$$

The last step is by the NISNID Fact, and by exercise **B1**(b) above. Using the strategy of the above example (and using part **B** above where necessary), complete the following exercises.

Exercise D1. Suppose Z is $N(-2,0.3)$. Find $P(-2.3 < Z < -1.7)$.

$$\begin{aligned} P(-2.3 < z < -1.7) &= P(\frac{-2.3-(-2)}{0.3} < \frac{z-(-2)}{0.3} < \frac{-1.7-(-2)}{0.3}) \\ &= P(-1 < \frac{z-(-2)}{0.3} < 1) = 0.683. \end{aligned}$$

Exercise D2. Suppose Z is $N(2,2)$. Find $P(-1.92 < Z < 5.92)$.

$$\begin{aligned} P(-1.92 < z < 5.92) &= P(\frac{-1.92-2}{2} < \frac{z-2}{2} < \frac{5.92-2}{2}) \\ &= P(-1.96 < \frac{z-2}{2} < 1.96) = 0.95. \end{aligned}$$

Exercise D3. Suppose Z is $N(\mu,\sigma)$. Find $P(\mu - 3\sigma < Z < \mu + 3\sigma)$.

$$\begin{aligned} P(\mu - 3\sigma < z < \mu + 3\sigma) &= P(\frac{\mu - 3\sigma - \mu}{\sigma} < \frac{z - \mu}{\sigma} < \frac{\mu + 3\sigma - \mu}{\sigma}) \\ &= P(-3 < \frac{z-\mu}{\sigma} < 3) = 0.997. \end{aligned}$$

Exercise D4. What exercise **D3** directly above says is: if Z is *any* normal random variable, then ___99.7___% of the data lies within three standard ___deviations___ of the ___mean___.

In this worksheet, we use calculus to derive some properties of the "standard normal curve." The worksheet is divided into two parts: Review (Part A), and Exercises (Part B).

Part A: Review. You may want to recall the following facts about **integration in polar coordinates** from Calculus II. **NOTE:** if you feel comfortable with integration in polar coordinates, you can skip this section, and go directly to **Part B: Exercises** below.

Suppose you have a region R in the xy plane, and you want to integrate some function $f(x, y)$ over that region. That is, suppose you want to compute the definite integral

$$I = \iint\limits_R f(x, y)\, dx\, dy. \qquad (*)$$

Sometimes, such an integral may be easier to do it you **switch to polar coordinates.** This means you replace x and y by their polar coordinate representations

$$x = r\cos\theta, \qquad y = r\sin\theta,$$

where r is the distance from the origin to the point (x, y), and θ is the angle (between 0 and 2π) that the point makes with the positive x-axis.

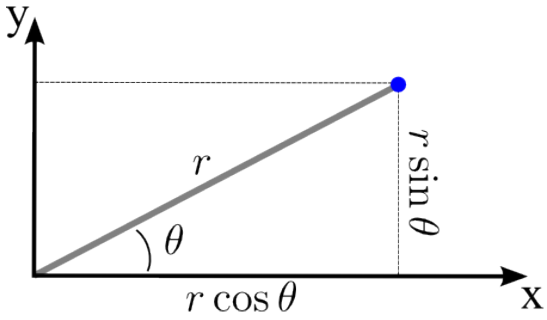


Figure 1. Polar coordinates

There are *two important things* to keep in mind when switching to polar coordinates:

1. You need to not only replace x and y by $r\cos\theta$ and $r\sin\theta$ in $f(x, y)$, but you also need to replace the "measure" $dx\, dy$ by $r\, dr\, d\theta$.

   In other words, we have

$$I = \iint\limits_R f(x, y)\, dx\, dy = \iint\limits_R f(r\cos\theta, r\sin\theta)\, r\, dr\, d\theta. \qquad (**)$$

2. You need to rewrite R in polar coordinates as well. That is: in the original formula $(*)$, R will typically be described in Cartesian (that is, xy) coordinates. To do the polar coordinate integral in $(**)$, you'll need to express R in polar coordinates.

OK, on to the Exercises.

Part B: Exercises. **Note:** you can do these exercises directly on this worksheet, or hand in your answers on your own (physical or virtual) paper. If you use separate paper, then for Exercise 1, which is a "fill in the blanks" exercise, you can just supply the words or formulas that go in the blanks. For Exercise 2, please show your work.

Exercise 1. As a warm-up, we'll evaluate the integral

$$\int_{x=0}^{1}\int_{y=0}^{\sqrt{1-x^2}} x\,(x^2+y^2)^4\,dx\,dy,$$

by switching to polar coordinates.

Solution. (Fill in the red blanks, there are ten of them.) First, let's work out the region of integration, in polar coordinates. We are integrating over the set

$$R = \{(x,y) : 0 \le x \le 1,\ 0 \le y \le \sqrt{1-x^2}\}.$$

It's clear what the condition $0 \le x \le 1$ means, but what about the condition $0 \le y \le \sqrt{1-x^2}$? This means we're looking at y values between the x axis (which is the line $y = 0$) and the curve $y = \underline{\quad\sqrt{1-x^2}\quad}$. But note that squaring both sides of $y = \sqrt{1-x^2}$ gives $y^2 = \left(\sqrt{1-x^2}\right)^2 = 1 - x^2$, so $x^2 + y^2 = \underline{\quad 1 \quad}$. The latter is the formula for the circle of radius one, centered at the point $\underline{\quad (0,0) \quad}$. So we're integrating over the region that's under that circle, over the y axis, and satisfying $0 \le x \le 1$. That is, we're integrating over this quarter-circle:

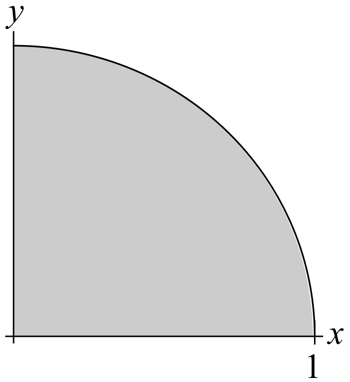


Figure 2. The region $R = \{(x,y) : 0 \le x \le 1,\ 0 \le y \le \sqrt{1-x^2}\}$

In polar coordinates, this region may be described as $R = \{(r,\theta) : 0 \le r \le 1,\ 0 \le \theta \le \frac{\pi}{2}\}$. Putting this information, together with the substitutions $x = r\cos\theta$, $y = \underline{\quad r\sin\theta \quad}$, and $dx\,dy = r\,dr\,d\theta$, into our integral in x and y, we find that

$$\int_{x=0}^{1}\int_{y=0}^{\sqrt{1-x^2}} x\left(x^2+y^2\right)^4 dx\,dy = \int_{r=0}^{1}\int_{\theta=0}^{\pi/2} r\cos\theta\left((r\cos\theta)^2+(r\sin\theta)^2\right)^4 r\,dr\,d\underline{\quad\theta\quad}.$$

Now, since $\cos^2\theta + \sin^2\theta = 1$ always, we find that

$$(r\cos\theta)^2 + (r\sin\theta)^2 = r^2(\cos^2\theta + \sin^2\theta) = \underline{\qquad r^2 \qquad},$$

so

$$\int_{r=0}^{1}\int_{\theta=0}^{\pi/2} r\cos\theta\left((r\cos\theta)^2 + (r\sin\theta)^2\right)^4 r\,dr\,d\theta = \int_{r=0}^{1}\int_{\theta=0}^{\pi/2} r\cos\theta\left(r^2\right)^4 r\,dr\,d\theta$$
$$= \int_{r=0}^{1}\int_{\theta=0}^{\pi/2} \cos\theta\, r^{10}\,dr\,d\theta.$$

Now note that the integral on the right breaks up into two separate integrals: one in the variable r, and the other in the variable θ. The first of these integrals equals

$$\int_{r=0}^{1} r^{10}\,dr = \left.\frac{r^{11}}{11}\right|_0^1 = \frac{1}{11}(1^{11} - 0^{11}) = \frac{1}{11}.$$

The second integral equals

$$\int_{\theta=0}^{\pi/2} \cos\theta\,d\theta = \sin\theta\Big|_0^{\pi/2} = \sin(\pi/2) - \sin(\underline{\qquad 0 \qquad}) = \underline{\qquad 1 \qquad}.$$

So the final result is

$$\int_{x=0}^{1}\int_{y=0}^{\sqrt{1-x^2}} x\left(x^2+y^2\right)^4 dx\,dy = \frac{1}{11} \times \underline{\qquad 1 \qquad} = \underline{\qquad \frac{1}{11} \qquad}.$$

Exercise 2. For this Exercise, answer parts (a)(b)(c)(d) below, and please show all your work. You can write your answers in the blank spaces provided on this worksheet, or on separate (physical or virtual) paper.

The purpose of this Exercise is to show that the standard normal curve

$$f_{0,1}(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2},$$

with domain equal to the entire real line, really IS a pdf. What does this mean? Well, recall that, for a pdf, probability equals area, so the area under a pdf must equal 1. So we want to show that this is true for $f_{0,1}(x)$. That is, we want to show that

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\,dx = 1,$$

or in other words, that

$$\int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \sqrt{2\pi}. \qquad (✸)$$

To do this, we're going to use a trick: we're going to multiply the integral by itself, to get a double integral, and then switch to polar coordinates.

(a) Let's call the integral that we're trying to evaluate, in (✸), I:

$$I = \int_{-\infty}^{\infty} e^{-x^2/2}\,dx.$$

Explain why

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy.$$ (🍂)

Hint: break the integral on the right up into a product of two integrals, using the fact that $e^{a+b} = e^a\,e^b$.

We have

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2/2}\,e^{-y^2/2}\,dx\,dy \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right) e^{-y^2/2}\,dy.\end{aligned}$$

Now as far as y is concerned, the integral inside the large parentheses above is just a constant, and we can pull constants outside of integrals. So we get

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy = \left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)\int_{-\infty}^{\infty} e^{-y^2/2}\,dy.$$

Both the integral in x and the integral in y, on the right hand side of the above, equal I, by definition of I. So we find that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy = I^2,$$

as claimed.

(b) Write the integral I^2 defined by (🍂) as an integral in polar coordinates. Hint: I^2 is an integral over the entire xy plane. In polar coordinates, this plane can be described as $\{(r,\theta) : 0 \le r < \infty,\ 0 \le \theta \le 2\pi\}$. **Don't forget that $dx\,dy$ becomes $r\,dr\,d\theta$!!** Also, you should simplify the quantity $(r\cos\theta)^2 + (r\sin\theta)^2$ that you get in your exponent, using the fact that $\cos^2\theta + \sin^2\theta = 1$.

You don't need to evaluate the (r,θ) integral that you get—yet.

We have

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy = \int_{r=0}^{\infty}\int_{\theta=0}^{2\pi} e^{-\left((r\cos\theta)^2+(r\sin\theta)^2\right)/2}\,r\,dr\,d\theta \\ &= \int_{r=0}^{\infty}\int_{\theta=0}^{2\pi} e^{-\left(r^2(\cos^2\theta+\sin^2\theta)\right)/2}\,r\,dr\,d\theta = \int_{r=0}^{\infty}\int_{\theta=0}^{2\pi} e^{-r^2/2}\,r\,dr\,d\theta. \end{aligned}$$

(c) If you did part **(b)** correctly, the integral that you get, in r and θ, should break up into an integral in r times an integral in θ. Evaluate each of these integrals using calculus (not using Wolfram Alpha, or a calculator, etc.). Please show all work. Some hints:

- **(i)** The integral in θ should be straightforward. Hint: you should get 2π for this integral.
- **(ii)** To do the integral in r, make the u-substitution $u = -r^2/2$. Hint: you should get 1 for this integral.

$$\int_{r=0}^{\infty}\int_{\theta=0}^{2\pi} e^{-r^2/2}\, r\, dr\, d\theta = \left(\int_{\theta=0}^{2\pi} d\theta\right)\left(\int_{r=0}^{\infty} e^{-r^2/2}\, r\, dr\right).$$

The integral in θ equals

$$\theta\Big|_0^{2\pi} = 2\pi - 0 = 2\pi.$$

To do the integral in r, we substitute $u = -r^2\, dr$. Then $du = -2r\, dr$. Also, when $r = 0$, $u = 0$, and when $r = \infty$, $u = -\infty$. So

$$\int_{r=0}^{\infty} e^{-r^2/2}\, r\, dr = \int_0^{-\infty} e^u\, (-du) = -\int_0^{-\infty} e^u\, du = -e^u\Big|_0^{-\infty} = -(e^{-\infty} - e^0) = -(0-1) = 1.$$

So, finally,

$$\int_{r=0}^{\infty}\int_{\theta=0}^{2\pi} e^{-r^2/2}\, r\, dr\, d\theta = \left(\int_{\theta=0}^{2\pi} d\theta\right)\left(\int_{r=0}^{\infty} e^{-r^2/2}\, r\, dr\right) = 2\pi \cdot 1 = 2\pi.$$

(d) To summarize your work on this problem, answer these questions: what is I^2 (as a number)? Use this information to answer: what is I? (Hint: I has to be positive, since it's the integral of a function that's always positive.) What is

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx \ ?$$

Finally, what is the area under the graph of $f_{0,1}$ (over its entire domain)? Note: some of the questions in this part may have the same answer.

We've just seen that $I^2 = 2\pi$. So $I = \pm\sqrt{2\pi}$. But as just noted, I is positive, so it must be that $I = \sqrt{2\pi}$. So the area under the graph of $f_{0,1}$ equals

$$\int_{-\infty}^{\infty} f_{0,1}(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} I = \frac{1}{\sqrt{2\pi}} \cdot \sqrt{2\pi} = 1,$$

so $f_{0,1}$ **is** a pdf.