Chapter 5: Continuous random variables

S520 online

These notes are written to accompany Trosset chapter 5.

Motivation: Random number generator

Trosset chapter 5.1.

Discrete random variables are intuitive: you list the outcomes and then you put probabilities on them. But they're not the only kind of random variable.

For example, suppose we create a random number generator that picks a real number X between 0 and 1. We could do it in the following way:

1. Roll a ten-sided die labeled $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$. This is the first digit after the decimal point.
2. Roll the die again. This is the second digit.
3. Roll the die again. This is the third digit.
4. Keep going for an infinite number of die rolls.

Now what's the probability that X is exactly 0.5? To get 0.5, you'd need to roll a 5, then a 0, then another 0, then another 0, and so ad infinitum. But the chance of rolling an infinite number of zeroes in a row is zero.

However, we can easily answer the question: What's the probability that $X \leq 0.5$? To be less than or equal to 0.5, either X is less than 0.5 or X equals 0.5. But we already said $P(X = 0.5) = 0$.

$$\begin{aligned}
P(X \leq 0.5) &= P(X < 0.5) + P(X = 0.5) \\
&= P(X < 0.5) + 0 \\
&= P(X < 0.5) \\
&= P(\text{first die is 0, 1, 2, 3, or 4}) \\
&= 0.5
\end{aligned}$$

In fact, for any y between 0 and 1, $P(X \leq y) = y$. We can write this down formally as a CDF:

$$F(y) = \begin{cases} 0 & y < 0 \\ y & 0 \leq y < 1 \\ 1 & y \geq 1 \end{cases}$$

This is a special case of a **uniform distribution**.

The CDF and the PDF

Trosset ch. 5.2 p. 120-123.

Continuous random variables have a continuous CDF. They have an uncountable number of possible outcomes.

The PMF fails <u>for continuous random variables</u> because the probability of any single number is zero. However, probability still works because the chance of being in an *interval* is not zero.

It turns out we can define something analagous to the PMF for the continuous case. A **probability density function**, or **PDF**, is a function $f(x)$ such that:

1. $f(x) \geq 0$ for all real numbers x.
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

To verify that a function is a PDF, check these two conditions above.

A continuous random variable X has a probability density function that allows you to find the probability of being in any interval by *integration*, i.e. by finding the area under the curve:

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

(Note that since the probability of any single number is zero, we could have also used "less than" in place of "less than or equal to" above.)

The above holds even when $a = -\infty$ or $b = \infty$.

The CDF is just the probability of being in the interval $(-\infty, y]$:

$$F(y) = \int_{-\infty}^{y} f(x)\,dx.$$

i.e., to get the CDF from PDF, integrate the PDF, $f(x)$, from $-\infty$ to a value of interest. In this course, integration is not required. Instead, integration of $f(x)$ from a to b is equivalent to the calculation of the area under $f(x)$ and above the horizontal line, between a and b.

To get the PDF from the CDF, differentiate.

Finally, note that:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

The probability something happens still has to be 1.

Expected value and variance

Trosset p. 123-124.

We define expected value and variance by analogy to the discrete case, only instead of taking sums, we take integrals.

Expected value:

$$\mu = EX = \int_{-\infty}^{\infty} x\,f(x)\,dx$$

Variance:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx = 1$$

Note: It's often easier to find the variance using the alternative formula

$$\mathrm{Var}(X) = EX^2 - (EX)^2$$

where

$$EX^2 = \int_{-\infty}^{\infty} x^2 \, f(x) \, dx$$

Note: The formula $EX^2 - (EX)^2$ also works in the discrete case.

Finally, the standard deviation σ is once again just the square root of the variance.

Example: Uniform random variables

In a **uniform random variable**, the probability of being in an interval is proportional to the length of that interval, as long as the interval is between a minimum a and a maximum b. We can write down this idea formally as a PDF:

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise.} \end{cases}$$

This is just a rectangle, scaled so that the total area under the rectangle is equal to 1.

To find the CDF, we integrate. Again, this is just equivalent to finding the area of a rectangle.

$$F(y) = \begin{cases} 0 & y < a \\ \frac{y-a}{b-a} & a \leq y < b \\ 1 & y \geq b \end{cases}$$

The expression to find the expected value is

$$EX = \int_a^b \frac{x}{b-a} \, dx$$

If you don't like calculus, we can find the expected value by thinking of it as the long-run average. Since the PDF is symmetric, the expected value must be at the point of symmetry, i.e. halfway between a and b. That is, $EX = (a+b)/2$.

To find the variance, there's no real alternative to doing the integration. We find that for the uniform, $\sigma^2 = (b-a)^2/12$.

Example. *The minute hand of my watch moves continuously until it stops at a random time. Let X be the position of the minute hand (in minutes after the hours) – it doesn't have to be a whole number. What are the expected value and variance of X?*

X is Uniform$(0, 60)$, so $EX = 30$ and $\text{Var}(X) = (60-0)^2/12 = 300$.

Example. *Let X be Uniform$(0, 1)$. Generate Y by first rolling a fair die:*

- *If the die shows 1 or 2, then $Y = X$.*
- *If the die shows 3, 4, 5, or 6, then $Y = X + 10$.*

What's the PDF of Y? What's the CDF? What's the expected value?

One-third of the time, Y is between 0 and 1, while 2/3rds of the time, it's between 10 and 11. The PDF is thus

$$f(x) = \begin{cases} \frac{1}{3} & 0 < x < 1 \\ \frac{2}{3} & 10 < x < 11 \\ 0 & \text{otherwise.} \end{cases}$$

The CDF is

$$
F(y) = \begin{cases}
0 & y < 0 \\
\frac{1}{3}y & 0 \le y < 1 \\
\frac{1}{3} & 1 \le y < 10 \\
\frac{1}{3} + \frac{2}{3}(y - 10) & y \le 10 < 11 \\
1 & y \ge 11
\end{cases}
$$

To find the expected value, you could either use integral or the property of expected value. Since the random variable Y is a Uniform $(0, 1)$ with probability $1/3$ and a Uniform$(10, 11)$ with probability $2/3$, it can be considered as a weighted average of two uniform random variables, i.e., $Y = 1/3Y_1 + 2/3Y_2$ where $Y_1 \sim Unif(0, 1)$ and $Y_2 \sim Unif(10, 11)$. Because of symmetry, $E(Y_1) = 0.5$, and $E(Y_2) = 10.5$, the expected value of Y is $E(Y) = 1/3E(Y_1) + 2/3E(Y_2) = (1/3 \times 0.5) + (2/3 \times 10.5) = 43/6$.

The Normal distribution

Trosset chapter 5.4

A Normal(μ, σ^2) random variable is a continuous random variable that has PDF

$$
f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2 \right]
$$

for *all* real values x. That is, there's no lowest or highest possible values (though values very far from the mean are extremely unlikely.)

A Normal(μ, σ^2) random variable has mean 0 and variance σ^2 (and hence standard deviation σ.) A **standard normal** random variable is a normal random variable with mean 0 and variance 1 (and SD 1.)

The R function `dnorm()` gives the PDF. We can graph the PDF of the standard normal:

```
curve(dnorm, from=-4, to=4)
abline(h = 0)
```
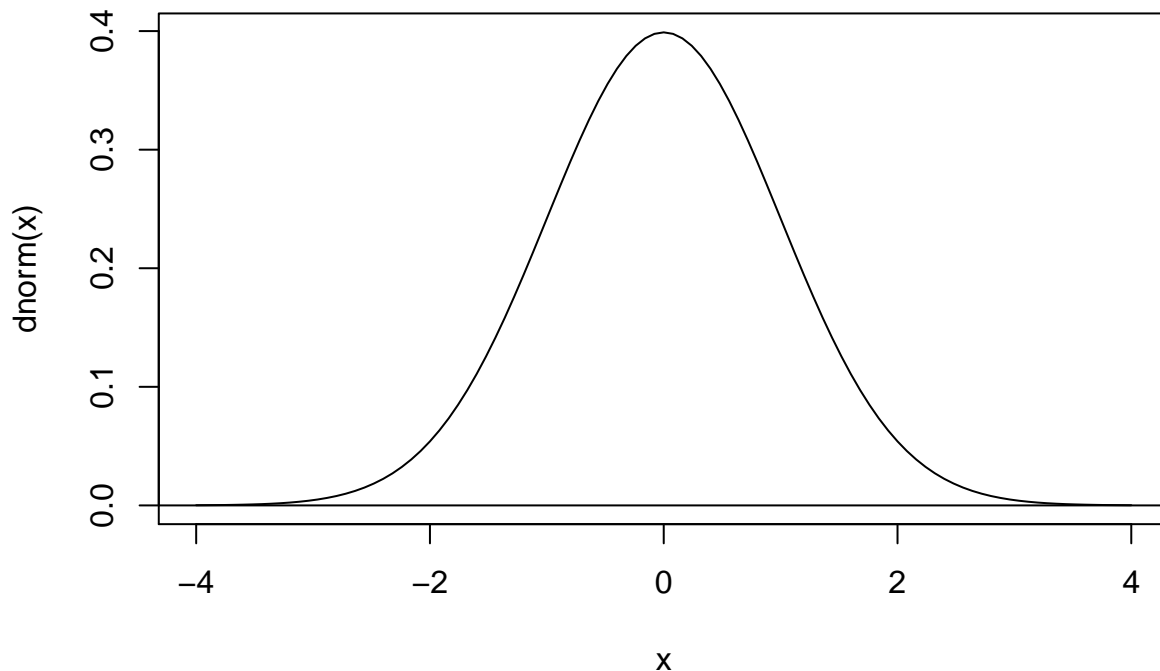
```
# Alternative code:
# x = seq(-4, 4, 0.01)
# f = dnorm(x)
# plot(x, f, type="l")
# abline(h=0)
```

Note how quickly the PDF dies down toward zero as we move away from the center of the distribution. All Normal distribution have this "bell shape": the mean parameter μ controls where the bell is centred, while the standard deviation σ controls how spread out the curve is. Here's the PDF of a Normal$(50, 5^2)$:

```
x = seq(30, 70, 0.01)
f = dnorm(x, mean=50, sd=5)
plot(x, f, type="l")
abline(h = 0)
```
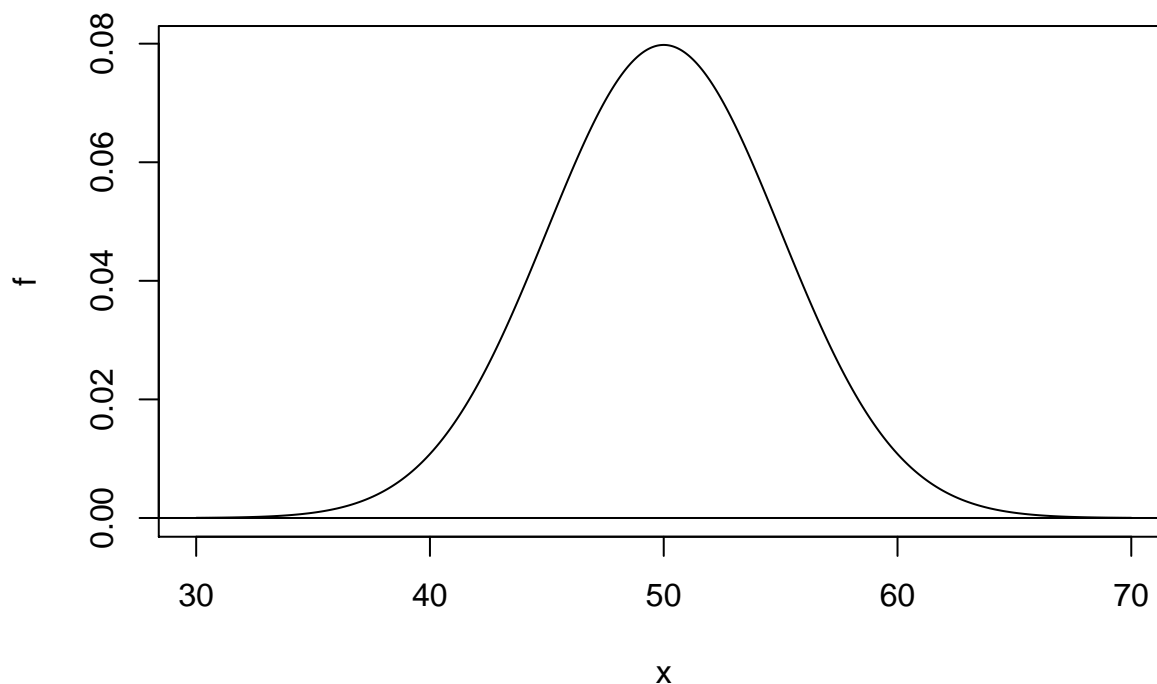

Note that very few things have truly Normal distributions. Instead, the Normal is often a good approximation to other distributions. For example, that Normal$(30, 5^2)$ is a good approximation to a binomial with $n = 100$ and $p = 0.5$:

```
data = rbinom(100000, 100, 0.5)
hist(data, prob=TRUE)
lines(x, f)
```
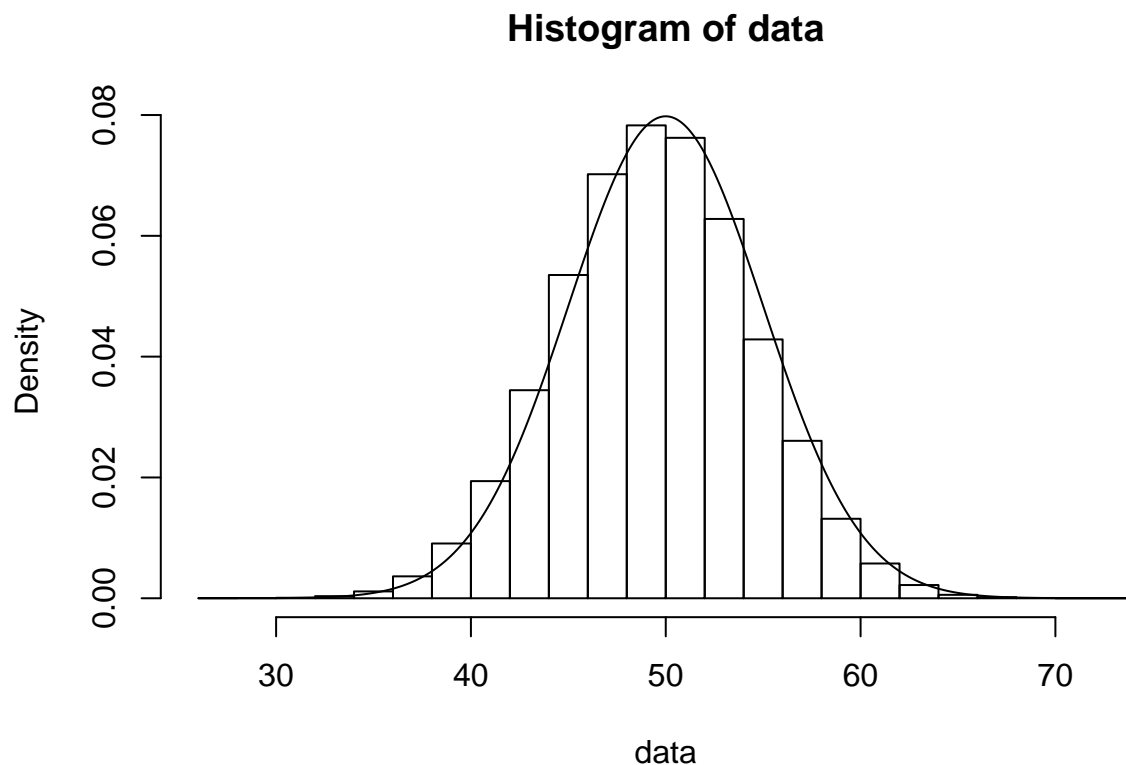
```

## Histogram of data



At this level, `dnorm()` is not especially useful. Instead, we want to be able to find Normal probabilities. Recall that the CDF of a continuous random variable is:

$$F(y) = \int_{-\infty}^{y} f(x)\, dx$$

Now $f(x)$ is a pain to integrate: there's no closed form solution, so it has to be done numerically. Fortunately, the R function `pnorm()` does this for us. `pnorm(y, m, s)` gives the CDF of a Normal$(m, s^2)$ at $y$, i.e. the probability the Normal$(m, s^2)$ random variable is less than (or equal to) $y$.

**pnorm() examples**

```
# Let Z have a std normal distribution
# P(Z <= 1) = P(Z < 1)
pnorm(1)
```

```
## [1] 0.8413447
```

```
# P(Z > 1) = P(Z >= 1)
1 - pnorm(1)
```

```
## [1] 0.1586553
```

```
# P(-1 < Z < 1) = P(|Z| < 1)
pnorm(1) - pnorm(-1)
```

```
## [1] 0.6826895
```

```
# P(-2 < Z < 2)
pnorm(2) - pnorm(-2)
```

```
## [1] 0.9544997
```

```
# P(-3 < Z < 3)
pnorm(3) - pnorm(-3)
```

```
## [1] 0.9973002
```

```
# P(|Z| > 1) = P(Z < -1) + P(Z > 1)
pnorm(-1) + (1 - pnorm(1))
```

```
## [1] 0.3173105
```

```
2 * pnorm(-1)
```

```
## [1] 0.3173105
```

```
2 * (1 - pnorm(1))
```

```
## [1] 0.3173105
```

```
# P(|Z| > 6)
2 * (1 - pnorm(6))
```

```
## [1] 1.973175e-09
```

```
# Let X be normal with mean 100, variance 10^2
# P(90 < X < 110)
pnorm(110, mean=100, sd=10) - pnorm(90, mean=100, sd=10)
```

```
## [1] 0.6826895
```

```
pnorm(110, 100, 10) - pnorm(90, 100, 10)
```

```
## [1] 0.6826895
```

```
# P(80 < X < 120)
pnorm(120, 100, 10) - pnorm(80, 100, 10)
```

```
## [1] 0.9544997
```

```
# P(70 < X < 130)
pnorm(130, 100, 10) - pnorm(70, 100, 10)
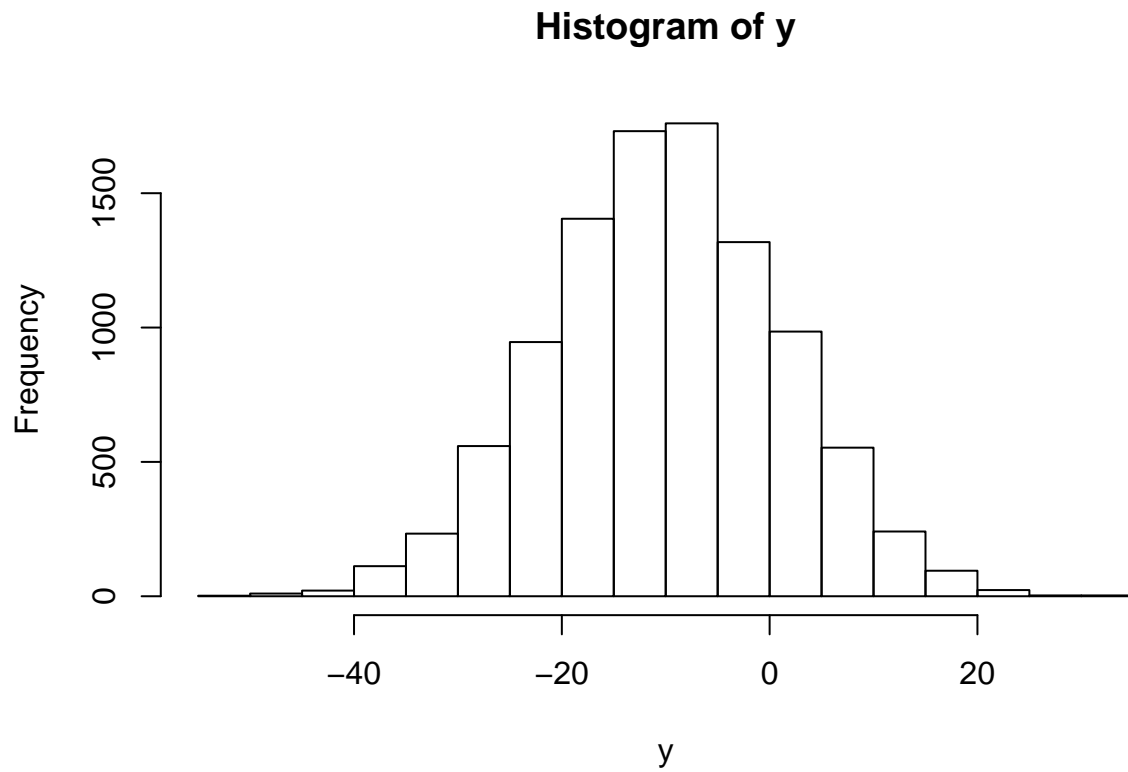```

```
## [1] 0.9973002
```

Combining Normal random variables

Adding, subtracting, or taking linear combinations of independent Normal random variables gives you a new normal random variable. The rules for finding the expected value and variance of the new random variable are the same as in the discrete case.

Example. *Let X_1 be Normal$(10, 5^2)$ and X_2 be Normal$(-20, 10^2)$, where X_1 and X_2 are independent. Let $Y = X_1 + X_2$. What's the distribution of Y?*

Y is Normal$(10 - 20, 5^2 + 10^2)$, i.e. Normal$(-10, 125)$.

```
x1 = rnorm(10000, mean=10, sd=5)
x2 = rnorm(10000, mean=-20, sd=10)
y = x1 + x2
hist(y)
```

Histogram of y

Note that multiplying two normals together does not result in a normal. (Read Trosset ch. 5.5 for other distributions we can derive from the normal; we'll skip this section for now.)