ITCS 121 Statistics

Lecture 7

Continuous Random Variables

Discrete random variables (review)

- In the last class we discussed some example of discrete random variables.
  - Binomial
  - Poisson
  - Hypergeometric
- Discrete random variables can only take whole number values.
- They are defined by probability mass functions, which give the probabilities for each possible value.
- For discrete random variables we have:

$$E(X) = \sum_{z \in \mathbb{Z}} zP(X = z) \qquad Var(X) = \sum_{z \in \mathbb{Z}} P(X = z)(z - E(X))^2$$

Continuous random variables

- Continuous random variables can take any real number value.
  - E.g. time to complete a puzzle, weight, temperature etc.
- A real number is any number that can be expressed as a (possibly infinitely long) decimal.
  - E.g. $1,\ {}^{1}/_{3},\ -5.667,\ 1.12\dot{3},\ \pi$ etc.
- Mathematicians often use the symbol $\mathbb{R}$ to represent the set of real numbers.
  - They also use $\mathbb{N}$ for the natural numbers, $\mathbb{Z}$ for the integers, and $\mathbb{Q}$ for the rational numbers (fractions).

Probability density functions

- Because a continuous random variable can take any real value, the probability that it takes any particular real value might always be zero.

- I.e. we could have $P(X = x) = 0$ for all x.

- So the probability mass function of a continuous random variable could be very uninformative.

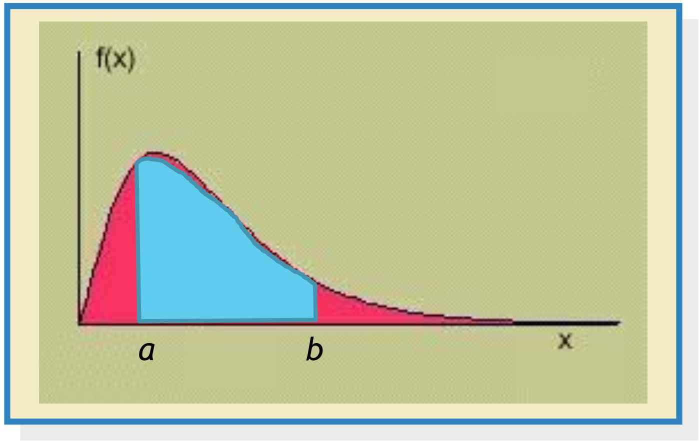
- To solve this problem, to describe continuous random variables where $P(X = x) = 0$ for all x we use something called a **probability density function** (pdf).

PDFs vs PMFs

- Let X be a continuous random variable and let f be its pdf.
- Similar to the pmf of a discrete random variable, the pdf f is a function from the set of real numbers to itself.
- But here $f(x)$ does *not* tell us the probability $P(X = x)$.
- The pdf does tell us how we can work out probabilities though.

Probabilities from PDFs

- Let X be a continuous random variable and let f be its pdf.
- We can draw the pdf f as a graph. E.g.



- Let a and b be real numbers. The probability that X takes a value between a and b is *the area under the curve between* $x = a$ *and* $x = b$.

Calculating probabilities

- How do we actually calculate $P(a < X < b)$ for a continuous random variable X with pdf f?
- On the previous slide we saw this is the area under the curve of f between $x = a$ and $x = b$.
- This is $\int_a^b f(x)\,dx$.
- So to find probabilities we have to integrate the pdf.
  - This might be hard because pdfs can be difficult functions.
- We can also work out things like $P(X < b)$ using $\int_{-\infty}^b f(x)\,dx$.

Some properties of PDFs.

- Notice that $P(-\infty < X < \infty)$ is the probability that X takes some real value, which has to be 1.
- It follows that $\int_{-\infty}^{\infty} f(x)\,dx = 1$ whenever f is the pdf of a random variable.

- Notice also that it doesn't matter when we integrate whether we include the endpoints or not.
- I.e. integrating over the open interval (a, b) is the same as integrating over the closed interval $[a, b]$.
- Translating back into probabilities, this means that for continuous random variables $P(X < a)$ is the same as $P(X \leq a)$, and so on.

Expected values for continuous random variables

- If X is a continuous random variable with pdf f, then we define its expected value using

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

- Actually calculating this integral can be hard.

- Like with discrete random variables, from the definition of the expected value we can get

$$Var(X) = E\left(\left(X - E(X)\right)^2\right) = E(X^2) - E(X)^2$$

The normal distribution

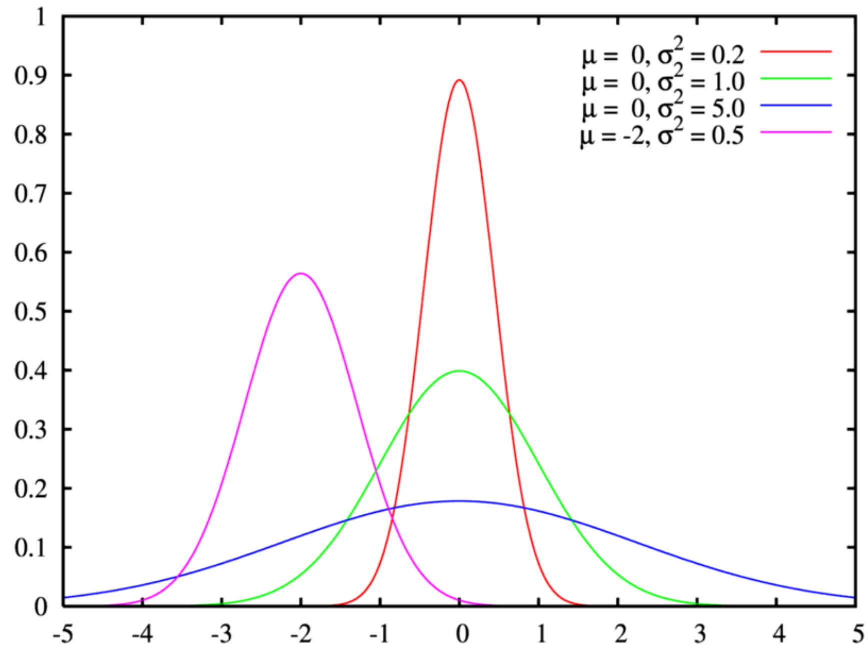
- A very useful distribution for continuous random variables is the **normal distribution.**
- A normal distribution has parameters μ (the mean) and σ (the standard deviation).
- It is described by a probability density function f as follows

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

- Here e is the important mathematical constant (like in the Poisson distribution), and has value ≈ 2.72.

What do normal distributions look like?

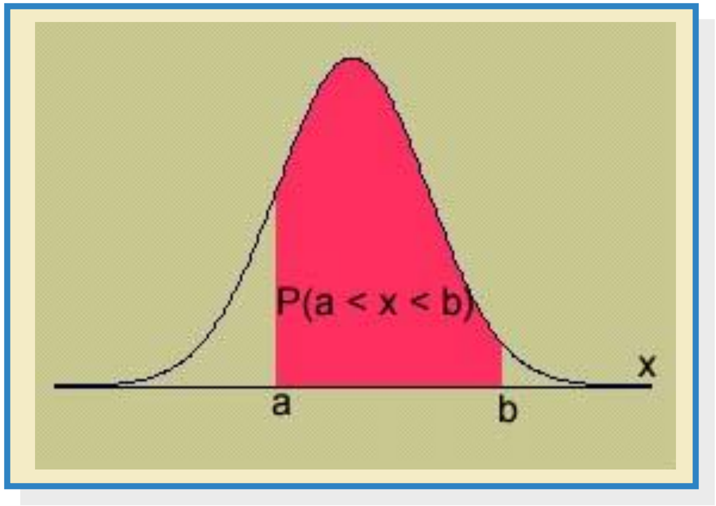
- The pdfs of normal distributions form nice symmetric, mound shaped graphs like this:
  - The mean μ determines where the peak of the graph is.
  - The standard deviation σ determines how tall or fat the graph is.



- If X is $Normal(\mu, \sigma)$ we can show that $E(X) = \mu$ and $Var(X) = \sigma^2$ (which is what we would expect!)

Finding normal probabilities

- If X is a $Normal(\mu, \sigma)$ random variable, then $P(a < X < b)$ is an area under a curve like this:



- We can, in theory, calculate it with $\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$.
- This is obviously not an easy integral to solve…

How we can actually find normal probabilities

- Fortunately we don't have to solve $\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \, dx$ every time we want to find a probability for a normal random variable.
- First we can simplify the integral a bit using the transformation $z = \frac{x-\mu}{\sigma}$ (remember this?).
- By some general calculus techniques we have

$$\int_{x=a}^{x=b} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \, dx = \int_{z=\frac{a-\mu}{\sigma}}^{z=\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \, dz$$

- The integral on the right is also not easy to solve, but fortunately for us other people have already solved it and put the results in a table, which we will see later.

The standard normal distribution

- The standard normal distribution is the normal distribution where $\mu = 0$ and $\sigma = 1$.
- If X is $Normal(\mu, \sigma)$ we can define new random variable Z by $Z = \frac{X-\mu}{\sigma}$.

$$\left.\begin{aligned} &E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{E(X)-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = 0. \\ &Var(Z) = Var\left(\frac{X-\mu}{\sigma}\right) = \frac{Var(X)}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1. \end{aligned}\right\} \text{So } Z \text{ is standard normal}$$

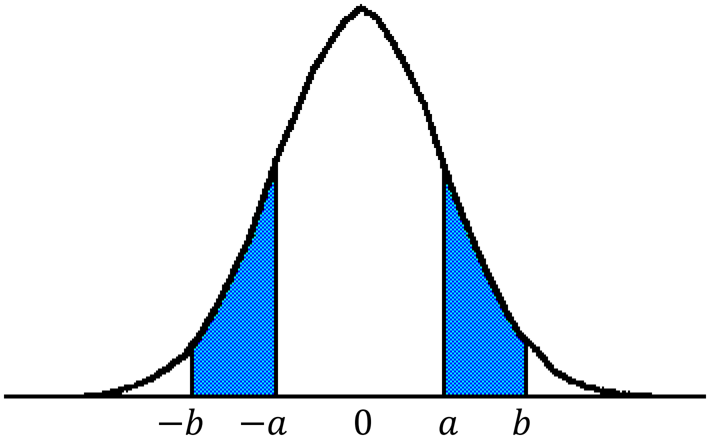
- The argument about integrals on the previous slide tells us that

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

- This is useful because we can use what we know about standard normal variables to find probabilities for general normal random variables

Properties of standard normal probabilities

- Standard normal distributions are symmetric around 0.
- This means $P(a < Z < b) = P(-b < Z < -a)$, for all values of a and b.



- We also have e.g. $P(Z > a) = 1 - P(Z < a)$, because $P(-\infty < Z < \infty) = 1$.

Example- standard normal probabilities

- Let Z be a standard normal random variable.
- Suppose we want to find $P(Z > 1.36)$.
- By looking at the standard normal distribution probability table, we see that $P(Z < 1.36) = 0.9131$.
- So $P(Z > 1.36) = 1 - 0.9131 = 0.0869$.
- Also, if we want to find $P(-1.2 < Z < 1.36)$ we can use

$$
\begin{aligned}
P(-1.2 < Z < 1.36) &= P(Z < 1.36) - P(Z < -1.2) \\
&= P(Z < 1.36) - (1 - P(Z < 1.2)) \\
&= 0.9131 - (1 - 0.8849) \\
&= 0.7980
\end{aligned}
$$

Class activity 1

Consider a standard normal random variable Z with mean $\mu = 0$ and standard deviation $\sigma = 1$. Use the table to find the following probabilities:

a) $P(Z < 2)$
b) $P(Z > 1.16)$
c) $P(-2.33 < Z < 2.33)$
d) $P(Z < 1.88)$

Example – normal probabilities

- The weights of packages of ground beef are normally distributed with mean 1 pound and standard deviation 0.10. What is the probability that a randomly selected package weighs between 0.80 and 0.85 pounds?
- Let X be the package weight. We want to find $P(0.8 < X < 0.85)$.
- So we want to find

$$P\left(\frac{0.8-1}{0.1} < Z < \frac{0.85-1}{0.1}\right) = P(-2 < Z < -1.5)$$

$$= P(1.5 < Z < 2)$$

$$= P(Z < 2) - P(Z < 1.5)$$

- According to the table this is $0.9772 - 0.9332 = 0.0440$.

Class activity 2

A normal random variable X has mean $\mu = 1.2$ and standard deviation $\sigma = 0.15$. Find the following probabilities.

a) $P(1.00 < X < 1.10)$

b) $P(X > 1.38)$

c) $P(1.35 < X < 1.50)$

The empirical rule again

- Remember the empirical rule:

Given a distribution of values that is approximately mound-shaped:

- Approximately 68% of the values occur within 1 standard deviation of the mean.
- Approximately 95% of the values occur within 2 standard deviations of the mean.
- Approximately 99.7% of the values occur within 3 standard deviations of the mean.

- If X is a normally distributed random variable then we have, for example

$$P(\mu-\sigma<X<\mu+\sigma)=P\left(\frac{(\mu-\sigma)-\mu}{\sigma}<Z<\frac{(\mu+\sigma)-\mu}{\sigma}\right)=P(-1<Z<1)$$

- Using the table we can calculate that $P(-1<Z<1)=0.6826$.
- We get similar results for $P(\mu-2\sigma<X<\mu+2\sigma)$ etc.
- In other words, the empirical rule comes from what we know about the normal distribution.

Working backwards

- We can work backwards from the standard normal distribution table to find z values giving a particular probability.
- E.g. If Z is a standard normal random variable, for what value of z do we have $P(Z < z) = 0.25$?

z	.00	.01	.02	.03	.04	.05	.06	.07	.08
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810

- According to the table this should be around $z = -0.675$.

Working backwards – another example

- Let X be a normal random variable with $\mu = 3$ and $\sigma = 2$. Find the value of x such that $P(X > x) = 0.05$.

- From the table, we see $P(Z > z) = 0.05$ when $z = 1.645$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

- Using $z = \frac{x-\mu}{\sigma}$ we see $x = 1.645 \times 2 + 3 = 6.29$.
- So $P(X > 6.29) = 0.05$.

Class activity 3

Let Z be a standard normal random variable.

a) Find z_0 such that $P(Z < z_0) = 0.9505$.

b) Find z_1 such that $P(-z_1 \leq Z \leq z_1) = 0.9$.

The normal approximation to the binomial random variable

- Remember the $Binomial(n, p)$ random variable which counts the number of successes in n independent trials where success has probability p.
- If X is a binomial random variable we saw we could use the Poisson distribution to approximate $P(X = k)$.
- We can also use the normal distribution to approximate $P(X \leq k)$.
- This can be much more convenient than calculating

$$P(X = 0) + P(X = 1) + \cdots + P(X = k)$$

When to use the normal approximation

- We can use the normal approximation when n is large and p is not too close to 0 or 1.
  - As a rough guide we sometimes demand $np > 5$ and $nq > 5$.

- A $Binomial(n, p)$ variable can be approximated as a $Normal(np, \sqrt{npq})$ variable.

- You'll have to wait till after the midterm to see an argument for why this is true, but we can apply this idea now.

Example – normal approximation

- Suppose X is $Binomial(30,0.4)$, and we want to find $P(X \leq 10)$.
- Because n is fairly large and p is close to ½ we can approximate X as a $Normal(np, \sqrt{npq})$ variable X'.
- That is, X' is $Normal(12, 2.683)$.
- So, we want to find $P(X \leq 10)$, and we are approximating the discrete variable X with the continuous variable X'.
- To get a discrete value out of continuous X' we have to round to the nearest integer.
- Since we are rounding like this we want to find $P(X' < 10.5)$.
  - This is known as a **continuity correction.**
- This translates to $P\left(Z < \frac{10.5-12}{2.683}\right) = P(Z < -0.56) = 0.2877$.
- With a little effort we can work out the true probability to be 0.2915 (to 4 decimal places).

Example - batteries

A production line produces AA batteries with a reliability rate of 95%. A sample of n = 200 batteries is selected. Find the probability that at least 195 of the batteries work.

> 5

Success = working battery, $n = 200$, $p = 0.95$, $np = 190$, $nq = 10$

We can use the normal approximation here.

Approximating Binomial X with normal X' we want to find $P(X \geq 195)$.

After the continuity correction this is $P(X' > 194.5)$.

This is $P\left(Z > \frac{194.5-np}{\sqrt{npq}}\right) = P\left(Z > \frac{194.5-190}{\sqrt{(200)(0.95)(0.05)}}\right) = P(Z > 1.46)$.

This is $1 - P(Z < 1.46)$, and we see from the table this is

$$1 - 0.9278 = 0.0722.$$

Class activity 4

Suppose that random variable X has a binomial distribution corresponding to n = 20 and p = 0.40. Use the normal approximation to the binomial probability distribution to calculate $P(X \geq 10)$.

Class activity 5

A normal random variable X has an unknown mean and standard deviation. The probability that X exceeds 4 is 0.9772, and the probability that X exceeds 5 is 0.9332. Find μ and σ.

Class activity 6

Suppose the numbers of a particular type of bacteria in samples of 1 ml of drinking water tend to be approximately normal distributed, with a mean of 85 and a standard deviation of 9. What is the probability that a given 1ml sample with contain more than 100 bacteria?