Continuous Probability Distributions

STA 198: Introduction to Health Data Science

Yue Jiang

June 03, 2020

Review: Discrete probability distributions

Event	Probability
$X = \textit{pre}$	0.10
$X = \textit{early}$	0.27
$X = \textit{full}$	0.57
$X = \textit{late/post}$	0.06

There are three rules for **discrete** probability distributions:

- Outcomes must be disjoint
- The probability of each outcome must be ≥ 0 and ≤ 1
- The sum of the outcome probabilities must add up to 1

Review: Expectation and variance

The expectation is the average value (weighted by the probability of each value occurring)

The variance describes the expected squared deviation of values from the population expectation

Can we be more precise?

Letting X be the random variable that corresponds to how long a baby's gestation was, we could imagine subdividing further and further:

Event	Prob.
$X < 20$ wk.	$P(X < 20)$
$X = 20$ to 21 wk.	etc.
$X = 21$ to 22 wk.	etc.
$X = 22$ to 23 wk.	etc.
$\vdots$	$\vdots$

Event	Prob.
$X < 20$ wk.	$P(X < 20)$
$X = 20$ to 20.1 wk.	etc.
$X = 20.1$ to 20.2 wk.	etc.
$X = 20.2$ to 20.3 wk.	etc.
$\vdots$	$\vdots$

Can we be more precise?

Now let gestational age X be a continuous random variable, which can take on *any* value, say from 0 to ∞. How might we define a continuous probability distribution that corresponds to X?

Continuous probability distributions

- The probability that a continuous variable equals any specific value is 0
- **No use tabulating** – there is an *uncountably* infinite number of possible values they can be, all with $P(X = x) = 0$
- The distribution is given by a probability density function, helps us describe probabilities for *ranges* of values

Density functions

Probability density functions satisfy the following two rules:

- The density must be non-negative everywhere ($f(x) \geq 0$ for all x from $-\infty$ to ∞)
    - This doesn't mean that it must range from $-\infty$ to ∞. We can have continuous distributions in a restricted range, for instance between (0, 1)
    - This only means that everywhere the density *is* defined, it is non-negative
- The total area under the density must be 1

Density functions

We can define events for continuous distributions and assign probabilities to them using density functions:

- Suppose X follows some density function $f(x)$
- We are interested in the event "X lies between a and b"
- We calculate the following probability:

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

(computers do this for us these days; no need to worry about the expression above)

What about other types of events?

Strict vs. non-strict inequalities

For continuous distributions, it does not matter whether we use strict or non-strict inequalities

$$
\begin{aligned}
P(a \leq X \leq b) &= P(X = a \; \cup \; a < X < b \; \cup \; X = b) \\
&= P(X = a) + P(a < X < b) + P(X = b) \\
&= P(a < X < b)
\end{aligned}
$$

The normal (Gaussian) distribution

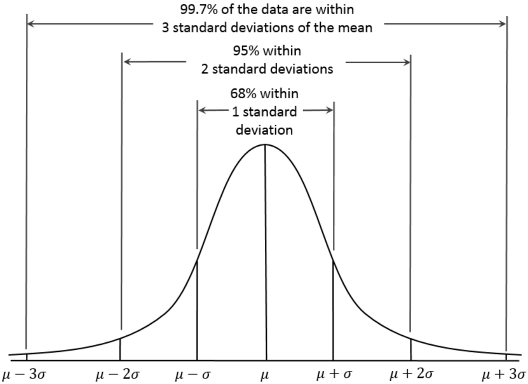
For the normal distribution,

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

where μ is the mean and σ^2 is the variance

We often write $N(\mu, \sigma^2)$

68-95-99.7


99.7% of the data are within
3 standard deviations of the mean
95% within
2 standard deviations
68% within
1 standard
deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Standardization

The normal distribution is a family of distributions of a specific form. There are an infinite amount of possible distributions, since μ can be any real number and σ^2 can be any positive number.

It would be very cumbersome to have to individually think about a $N(0, 20)$ vs. $N(2.5, 2)$ vs. $N(694, 1549)$ vs. .... distribution, depending on the situation

In practice, we could calculate a standard score that gives the number of standard deviations away from the mean an observation from a particular population is.

Why would we want to standardize?

z-scores

A z-score tells us how many population standard deviations an observation is away from the population mean

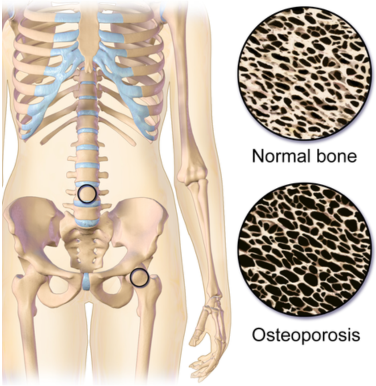
They provide ways to compare results across many different measurement scales, since z-scores are *unitless*

$$z = \frac{x - \mu}{\sigma}$$

(note the use of population parameters μ and σ)

So, a z-score of 1.2 is 1.2 standard deviations above the mean; a z-score of -0.8 is 0.8 standard deviations below the mean

Osteoporosis



According to NHANES, the mean bone mineral density for a 65 year old white woman is 809 mg/cm^3, with a standard deviation of 140 mg/cm^3.

Suppose you are a 65 year old white woman whose bone density is 698 mg/cm^3.

Are you very concerned about osteoporosis?