STATS8: Introduction to Biostatistics

Random Variables and Probability Distributions

Babak Shahbaba
UCI, Spring of 2012

Random variables

- In this lecture, we will discuss random variables and their probability distributions.

- Formally, a **random variable** X assigns a numerical value to each possible outcome (and event) of a random phenomenon.

- For instance, we can define X based on possible genotypes of a bi-allelic gene **A** as follows:

$$X = \begin{cases} 0 & \text{for genotype } AA, \\ 1 & \text{for genotype } Aa, \\ 2 & \text{for genotype } aa. \end{cases}$$

- Alternatively, we can define a random, Y, variable this way:

$$Y = \begin{cases} 0 & \text{for genotypes } AA \text{ and } aa, \\ 1 & \text{for genotype } Aa. \end{cases}$$

Random variables

- After we define a random variable, we can find the probabilities for its possible values based on the probabilities for its underlying random phenomenon.

- This way, instead of talking about the probabilities for different outcomes and events, we can talk about the probability of different values for a random variable.

- For example, suppose $P(AA) = 0.49$, $P(Aa) = 0.42$, and $P(aa) = 0.09$.

- Then, we can say that $P(X = 0) = 0.49$, i.e., X is equal to 0 with probability of 0.49.

- Note that the total probability for the random variable is still 1.

Random variables

- The probability distribution of a random variable specifies its possible values (i.e., its range) and their corresponding probabilities.
- For the random variable X defined based on genotypes, the probability distribution can be simply specified as follows:

$$P(X = x) = \begin{cases} 0.49 & \text{for } x = 0, \\ 0.42 & \text{for } x = 1, \\ 0.09 & \text{for } x = 2. \end{cases}$$

  Here, x denotes a specific value (i.e., 0, 1, or 2) of the random variable.

Discrete vs. continuous random variables

- We divide the random variables into two major groups: **discrete** and **continuous**.
- Discrete random variables can take a countable set of values.
- These variables can be categorical (nominal or ordinal), such as genotype, or counts, such as the number of patients visiting an emergency room per day,
- Continuous random variables can take an uncountable number of possible values.
- For any two possible values of this random variable, we can always find another value between them.

Probability distribution

- The probability distribution of a random variable provides the required information to find the probability of its possible values.
- The probability distributions discussed here are characterized by one or more **parameters**.
- The parameters of probability distributions we assume for random variables are usually unknown.
- Typically, we use Greek alphabets such as μ and σ to denote these parameters and distinguish them from known values.
- We usually use μ to denote the mean of a random variable and use σ^2 to denote its variance.

Discrete probability distributions

- For discrete random variables, the probability distribution is fully defined by the **probability mass function (pmf)**.

- This is a function that specifies the probability of each possible value within range of random variable.

- For the genotype example, the pmf of the random variable X is

$$P(X = x) = \begin{cases} 0.49 & \text{for } x = 0, \\ 0.42 & \text{for } x = 1, \\ 0.09 & \text{for } x = 2. \end{cases}$$

Bernoulli distribution

- Binary random variables (e.g., healthy/diseased) are abundant in scientific studies.

- The binary random variable X with possible values 0 and 1 has a **Bernoulli** distribution with parameter θ.

- Here, $P(X = 1) = \theta$ and $P(X = 0) = 1 - \theta$.

- For example,

$$P(X = x) = \begin{cases} 0.2 & \text{for } x = 0, \\ 0.8 & \text{for } x = 1. \end{cases}$$

- We denote this as $X \sim \mathrm{Bernoulli}(\theta)$, where $0 \leq \theta \leq 1$.

Bernoulli distribution

- The mean of a binary random variable, X, with Bernoulli(θ) distribution is θ. We show this as $\mu = \theta$.

- The variance of a random variable with Bernoulli(θ) distribution is $\sigma^2 = \theta(1-\theta) = \mu(1-\mu)$.

- The standard deviation is obtained by taking the square root of variance: $\sigma = \sqrt{\theta(1-\theta)} = \sqrt{\mu(1-\mu)}$.

Binomial distribution

- A sequence of binary random variables $X_1, X_2, \ldots, X_n$ is called **Bernoulli trials** if they all have the same Bernoulli distribution and are independent.

- The random variable Y representing the number of times the outcome of interest occurs in n Bernoulli trials (i.e., the sum of Bernoulli trials) has a $\mathrm{Binomial}(n, \theta)$ distribution.

- The pmf of a binomial(n, θ) specifies the probability of each possible value (integers from 0 through n) of the random variable.

- The theoretical (population) mean of a random variable Y with Binomial(n, θ) distribution is $\mu = n\theta$. The theoretical (population) variance of Y is $\sigma^2 = n\theta(1 - \theta)$.

Continuous probability distributions

- For discrete random variables, the pmf provides the probability of each possible value.

- For continuous random variables, the number of possible values is uncountable, and the probability of any specific value is zero.

- For these variables, we are interested in the probability that the value of the random variable is within a specific interval from x_1 to x_2; we show this probability as $P(x_1 < X \leq x_2)$.

Probability density function

- For continuous random variables, we use **probability density functions** (pdf) to specify the distribution. Using the pdf, we can obtain the probability of any interval.

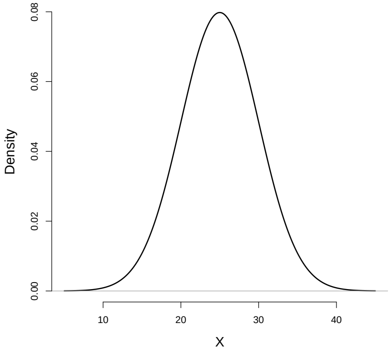
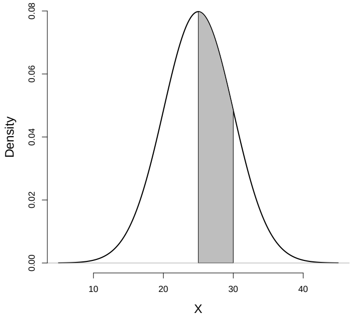


Figure: Probability density function for BMI

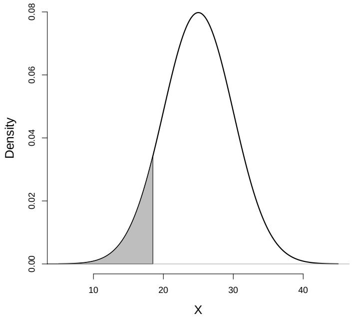
Probability density function

- The total area under the probability density curve is 1.
- The curve (and its corresponding function) gives the probability of the random variable falling within an interval.
- This probability is equal to the area under the probability density curve over the interval.



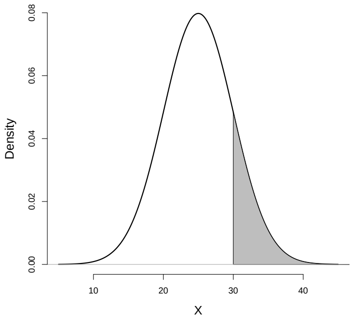
Lower tail probability

- the probability of observing values less than or equal to a specific value x, is called the lower tail probability and is denoted as $P(X \leq x)$.



Upper tail probability

- The probability of observing values greater than x, $P(X > x)$, is called the upper tail probability and is found by measuring the area under the curve to the right of x.



Probability of intervals

- The probability of any interval from x_1 to x_2, where $x_1 < x_2$, can be obtained using the corresponding lower tail probabilities for these two points as follows:

$$P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1).$$

- For example, the probability of a BMI between 25 and 30 is

$$P(25 < X \leq 30) = P(X \leq 30) - P(X \leq 25).$$

Normal distribution

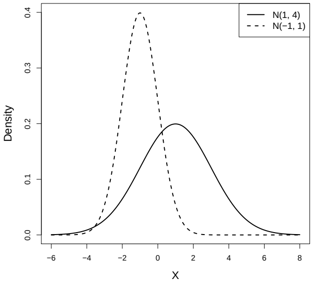
- Consider the probability distribution function and its corresponding probability density curve we assumed for BMI in the above example.

- This distribution is known as **normal** distribution, which is one of the most widely used distributions for continuous random variables.

- Random variables with this distribution (or very close to it) occur often in nature.

Normal distribution

- A **normal distribution** and its corresponding pdf are fully specified by the mean μ and variance σ^2.
- A random variable X with normal distribution is denoted $X \sim N(\mu, \sigma^2)$.
- $N(0, 1)$ is called the *standard normal distribution*.



The 68-95-99.7% rule

- The 68–95–99.7% rule for normal distributions specifies that
  - 68% of values fall within 1 standard deviation of the mean:

    $$P(\mu - \sigma < X \leq \mu + \sigma) = 0.68.$$

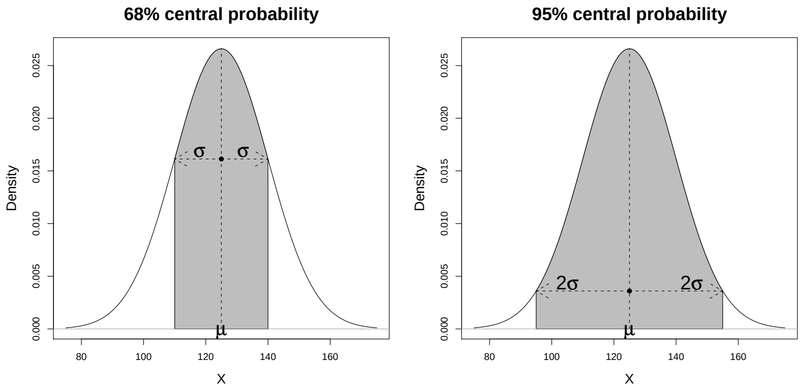
  - 95% of values fall within 2 standard deviations of the mean:

    $$P(\mu - 2\sigma < X \leq \mu + 2\sigma) = 0.95.$$

  - 99.7% of values fall within 3 standard deviations of the mean:

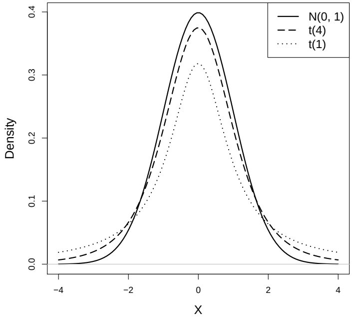
    $$P(\mu - 3\sigma < X \leq \mu + 3\sigma) = 0.997.$$

Normal distribution



Student's t-distribution

- Another continuous probability distribution that is used very often in statistics is the **Student's *t*-distribution** or simply the ***t*-distribution**.



Student's t-distribution

- A t-distribution is specified by only one parameter called the **degrees of freedom** *df*.

- The t-distribution with *df* degrees of freedom is usually denoted as $t(df)$ or t_{df}, where df is a positive real number $(df > 0)$.

- The mean of this distribution is $\mu = 0$, and the variance is determined by the degrees of freedom parameter, $\sigma^2 = df/(df - 2)$, which is of course defined when $df > 2$.

Cumulative distribution function

- We saw that by using lower tail probabilities, we can find the probability of any given interval.
- Indeed, all we need to find the probabilities of any interval is a function that returns the lower tail probability at any given value of the random variable: $P(X \leq x)$.
- This function is called the **cumulative distribution function** (cdf) or simply the **distribution function**.

Quantiles

- We can use the cdf plot in the reverse direction to find the value of the random variable for a given lower tail probability.

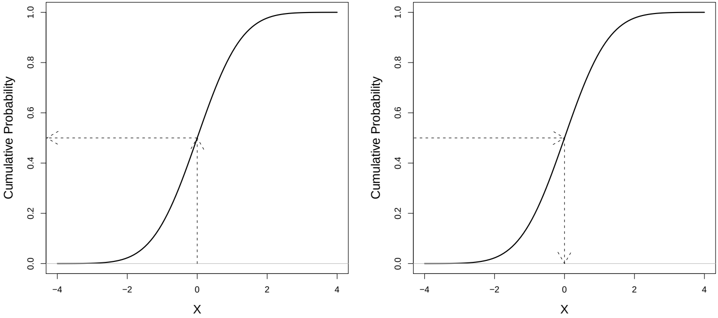


Figure: Left: Finding lower tail probabilities. Right: Finding quantiles

Scaling and shifting random variables

- If $Y = aX + b$, then

$$\begin{aligned} \mu_Y &= a\mu_X + b, \\ \sigma_Y^2 &= a^2\sigma_X^2, \\ \sigma_Y &= |a|\sigma_X. \end{aligned}$$

- The process of shifting and scaling a random variable to create a new random variable with mean zero and variance one is called **standardization**.

- For this, we first subtract the mean μ and then divide the result by the standard deviation σ.

$$Z = \frac{X - \mu}{\sigma}.$$

- If $X \sim N(\mu, \sigma^2)$, then $Z \sim N(0, 1)$.

Adding/subtracting random variables

- If $W = X + Y$, then
$$\mu_W = \mu_X + \mu_Y.$$
- If the random variables X and Y are independent (i.e., they do not affect each other probabilities), then we can find the variance of W as follows:
$$\sigma_W^2 = \sigma_X^2 + \sigma_Y^2.$$
- If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then assuming that the two random variables are independent, we have
$$W = X + Y \sim N\left(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\right).$$

Adding/subtracting random variables

- If we subtract Y from X, then

$$\mu_W = \mu_X - \mu_Y.$$

- If the two variables are independent,

$$\sigma_W^2 = \sigma_X^2 + \sigma_Y^2.$$

- Note that we still *add* the variances.
- Subtracting Y from X is the same as adding $-Y$ to X.