

P8130: Biostatistical Methods I

Lecture 5: Continuous Probability Distributions

Cody Chiuzan, PhD

Department of Biostatistics

Mailman School of Public Health (MSPH)

Lecture 4: Recap

- Randomness and random variables
- Binomial distribution: definition and statistical properties
- Poisson distribution: definition and statistical properties

Lecture 5: Outline

- Continuous random variables and probability distributions
- Uniform distribution: definition and statistical properties
- Normal distribution: definition and statistical properties

Continuous Random Variables

Compared to discrete r. v., continuous random variables can assume any value over an entire interval.

Continuous distributions are typically represented by a probability density function (pdf) or density curve.

A density curve is a representation of the underlying population distribution (not a description of the sample data)

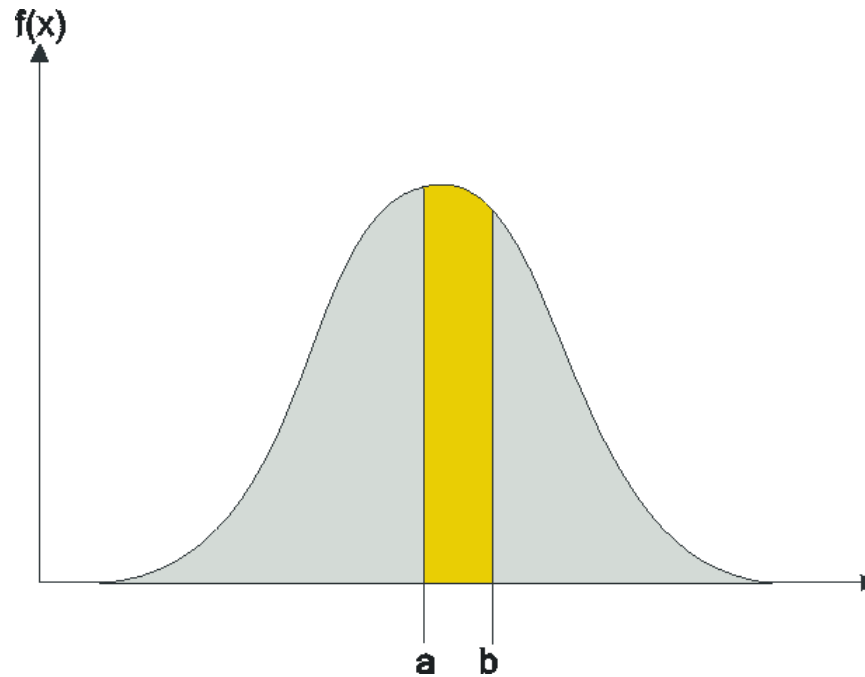
Properties of Density Functions (PDF)

The *pdf* of a continuous random variable X is a function defined for all real numbers x such that:

1. $f_X(x) = f(x) \geq 0$, for all values of x
2. The density function is always above or on the horizontal axis (curve cannot have a negative value)
3. The total region/area between the curve and horizontal axis is exactly 1.
4. For any real numbers a and b , $P(a \leq X \leq b)$ is given by the area bounded by the graph of f , the vertical lines $x = a$ and $x = b$, and the x axis.

Properties of Density Functions (PDF)

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



Also, all the following probabilities are equal:

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

Cumulative Distribution Function

The cumulative distribution of a continuous r. v. is denoted by:

$$F(x) = P(X \leq t) = \int_{-\infty}^t f(x)dx, -\infty < t < \infty$$

Properties:

1. $F(x)$ is a non-decreasing function, $t_1 \leq t_2$ implies that $F(t_1) \leq F(t_2)$
2. $F(x)$ is continuous
 - For discrete random variables, $F(x)$ is only right continuous
3. $F(x)' = f(x)$, if $F(x)'$ exists
4. $P(X > a) = 1 - F(a)$

$$P(a < X < b) = F(b) - F(a)$$

Expected Value of a Continuous Random Variable

The expected value of a continuous random variable is defined as:

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

The variance of a continuous random variable is given by:

$$\sigma^2 = \text{var}(X) = E(x^2) - \mu^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx$$

PDF: Example

Let the continuous random variable X have the following PDF:

$$f(x) = 4x^3, 0 \leq x \leq 1$$

Let us compute the following:

1. $\int_0^1 f(x)dx =$

2. $P(0.2 < X < 0.5) =$

3. $E(X) = \int_0^1 xf(x)dx =$

4. $var(x) = \int_0^1 x^2 f(x)dx - [E(x)]^2 =$

Uniform Distribution

The probability density function of a uniform r. v. X , $X \sim \text{Unif}(a, b)$, is given by:

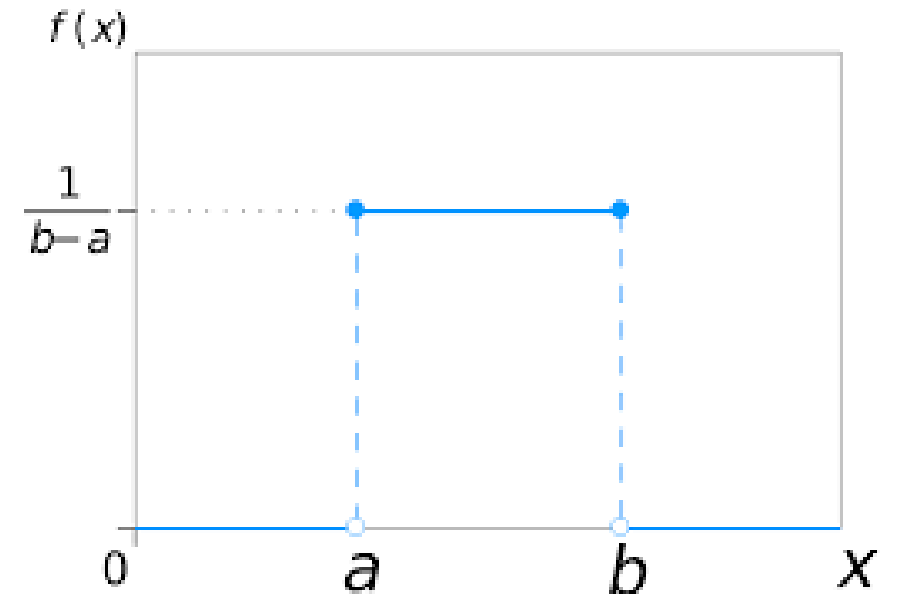
$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$$

The expected value of the uniform distribution is given by:

$$\mu = E(X) = \int_a^b x f(x) dx = \frac{b+a}{2}$$

The variance of the uniform distribution is given by:

$$\sigma^2 = \text{var}(X) = \int_a^b x^2 f(x) dx - \mu^2 = \frac{(b-a)^2}{12}$$



Uniform Distribution: Examples

An operator just announced a 30min delay for your subway. What is the probability that the subway will arrive between 15 and 20 min?

$$X \sim \text{Unif}(0,30)$$

The probability density function is given by:

$$f(x) = \frac{1}{30-0} = \frac{1}{30}, 0 < x < 30$$

$$P(15 < x < 20) = \int_{15}^{20} f(x) dx = \dots = \frac{5}{30} = \frac{1}{6}$$

↑
In class derivation

Normal Distribution

Probably the most common distribution used in statistics, the normal distribution is also called the *Gaussian* or the '*bell-shaped*' distribution.

- > Some variables are normal
- > Some are approximately normal
- > Some can be transformed to be approximately normal
- > The sampling error of the means tends towards normality even for non-normal populations

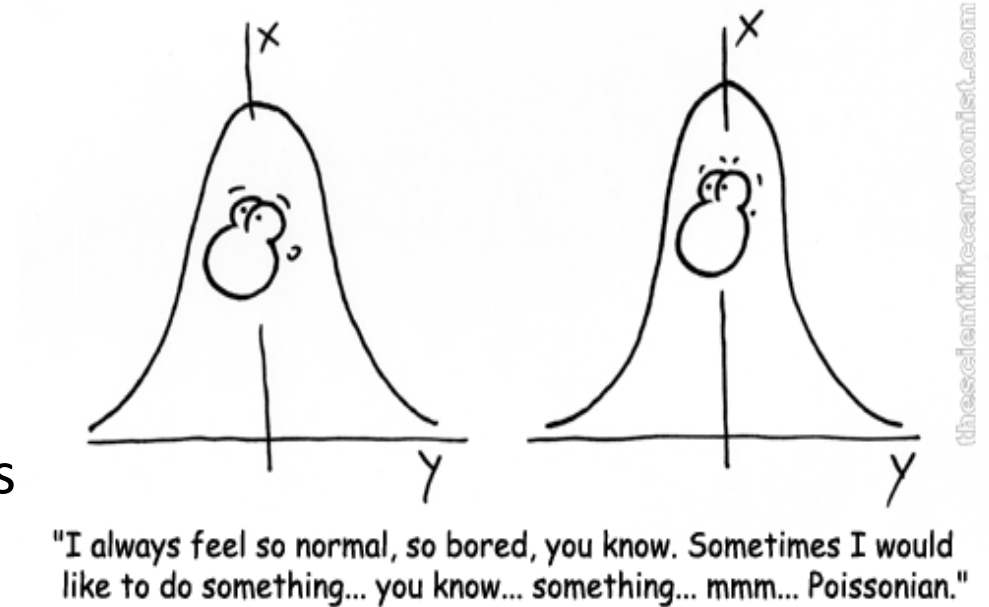


Image taken from the Scientific Cartoonist

Normal Distribution

The probability distribution function of a normal r. v. X , $X \sim N(\mu, \sigma^2)$, is given by:

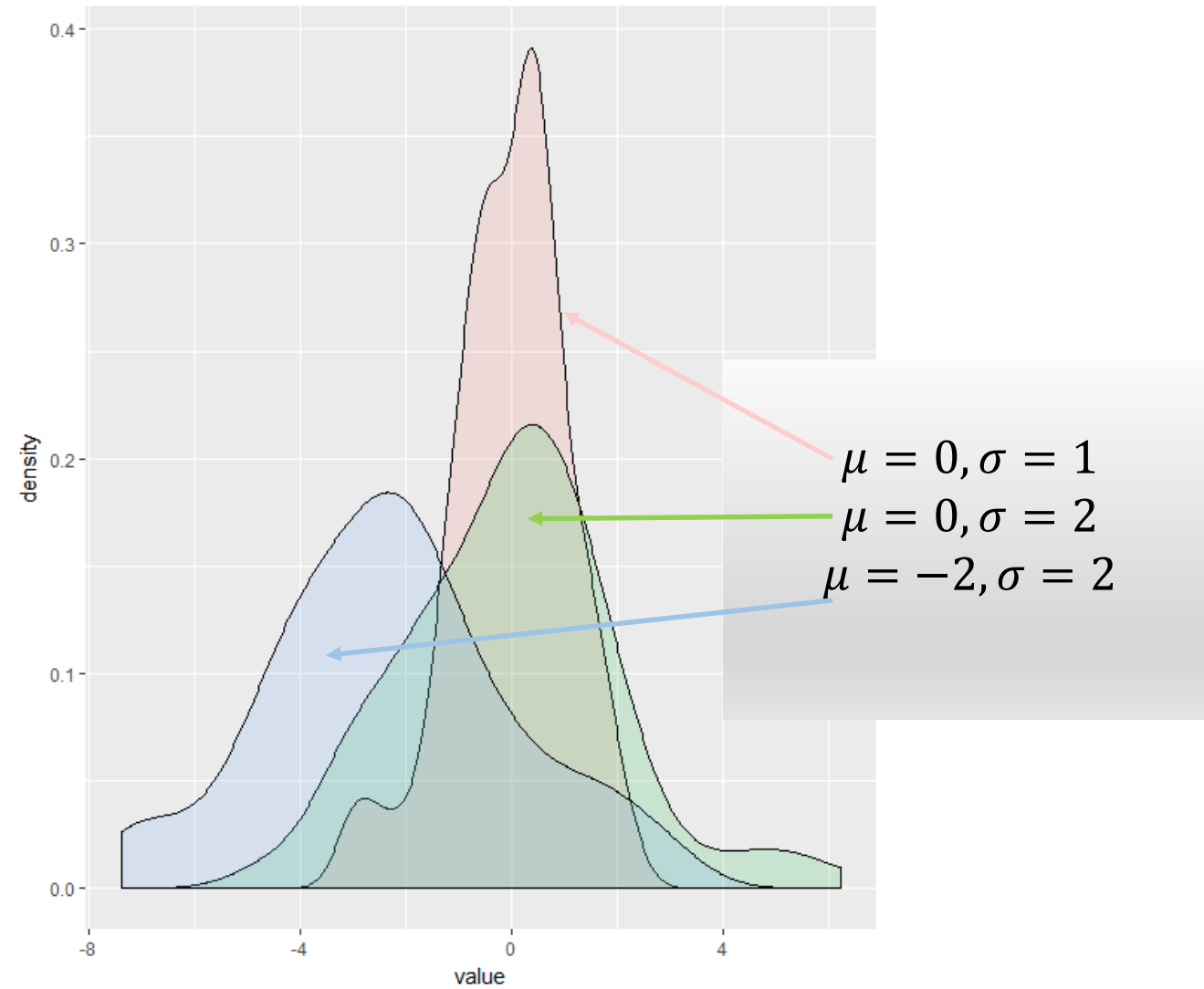
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty,$$

With parameters μ and σ , where $-\infty < \mu < \infty$, and $\sigma > 0$.

Properties:

1. The mean μ describes the center of the distribution
2. The standard deviation σ describes how much the curve is spread around the center
3. The normal distribution is symmetric around the mean

Normal Distribution



Standard Normal Distribution

A normal distribution with mean 0 and variance 1 is referred to as a 'standard normal' or 'unit normal distribution'.

The probability density function of a standard normal denoted by $N(0,1)$, is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, -\infty < x < \infty \quad (1)$$

If $X \sim N(\mu, \sigma)$ then $Z = \frac{X - \mu}{\sigma} \sim N(0,1)$. Thus, an alternative notation for (1) is:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}, -\infty < z < \infty \quad (2)$$

Why do we even need the standard normal distribution?

Standard Normal Distribution: Empirical Rule

Based on the Empirical Rule, for the standard normal distribution:

$$P(-1 < z < 1) = 0.68$$

$$P(-2 < z < 2) = 0.95$$

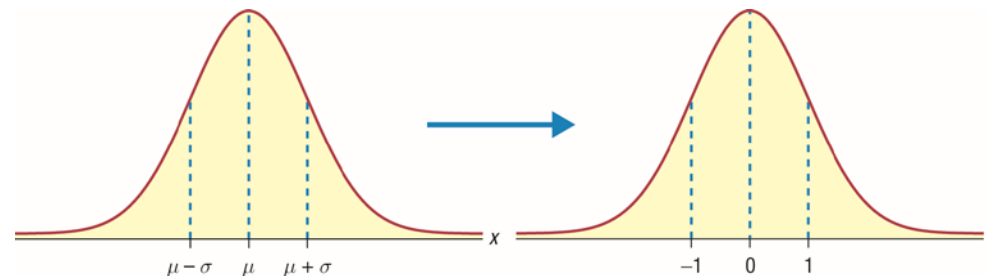
$$P(-3 < z < 3) = 0.99$$

This rule applies to any *approximately* bell-shaped distribution and it is usually interpreted as:

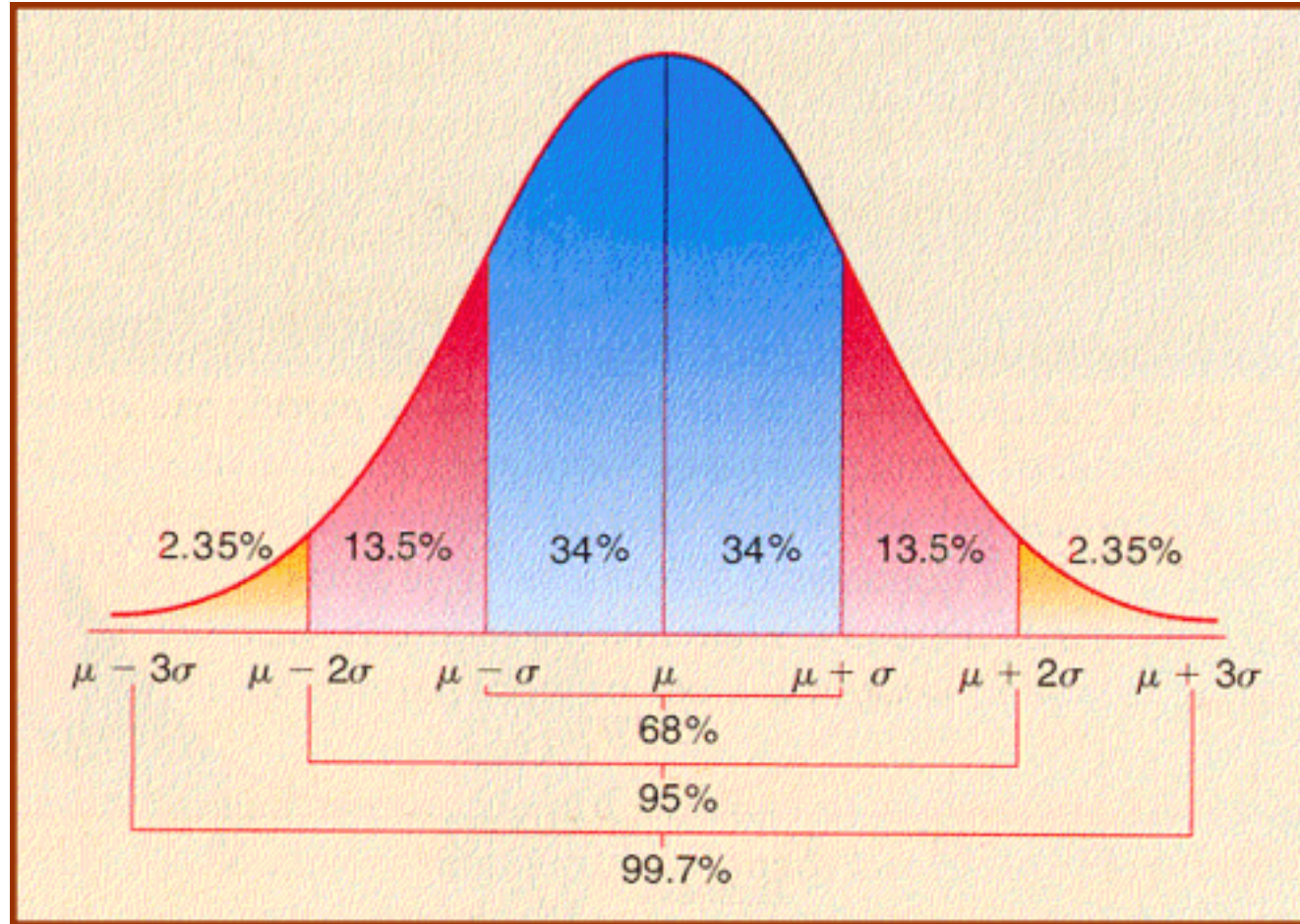
Approximately 68% of all values fall within one standard deviation from the mean.

Approximately 95% of all values fall within two standard deviations from the mean.

Approximately 99% of all values fall within three standard deviations from the mean.



Standard Normal Distribution: Empirical Rule



Standard Normal: Cumulative Distribution Function

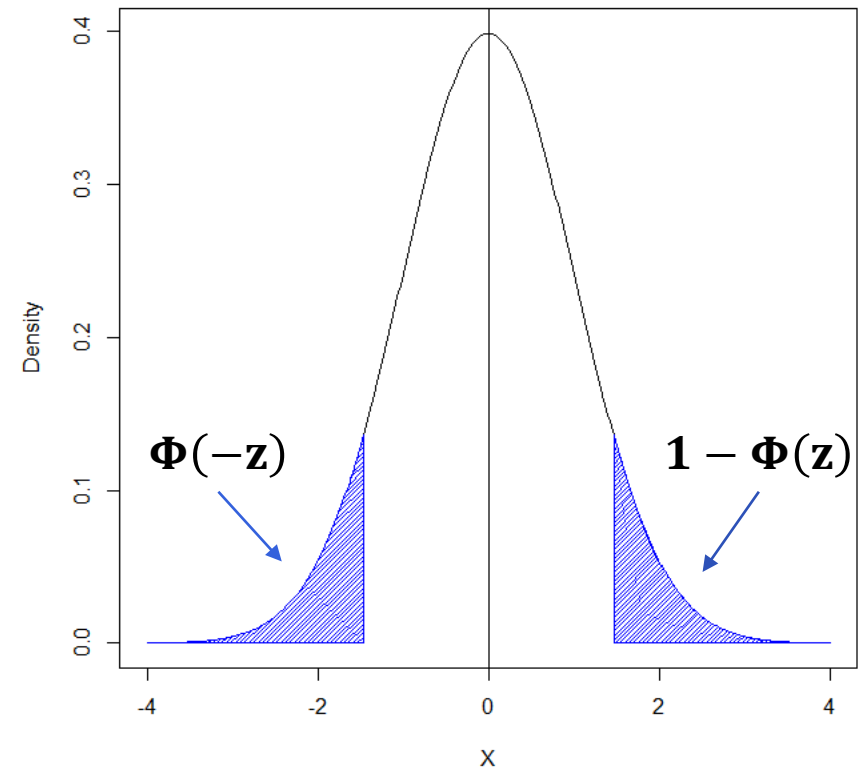
The cumulative distribution function (CDF) of a standard normal is denoted by:

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^z f(z)dz, -\infty < z < \infty$$

From the symmetry of standard normal:

$$\Phi(-z) = P(Z \leq -z) = P(Z \geq z)$$

$$P(Z \geq z) = 1 - P(Z \leq z) = 1 - \Phi(z)$$



Z-transformation

It is used to transform any normal r. v. to a standard normal r. v. We have already stated that:

$$\text{If } X \sim N(\mu, \sigma) \text{ then } Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

The 'z-score' measures how many standard deviations your observation is from the mean.

Example: Let X be a variable normally distributed with mean $\mu = 70$ and standard deviation $\sigma = 10$.

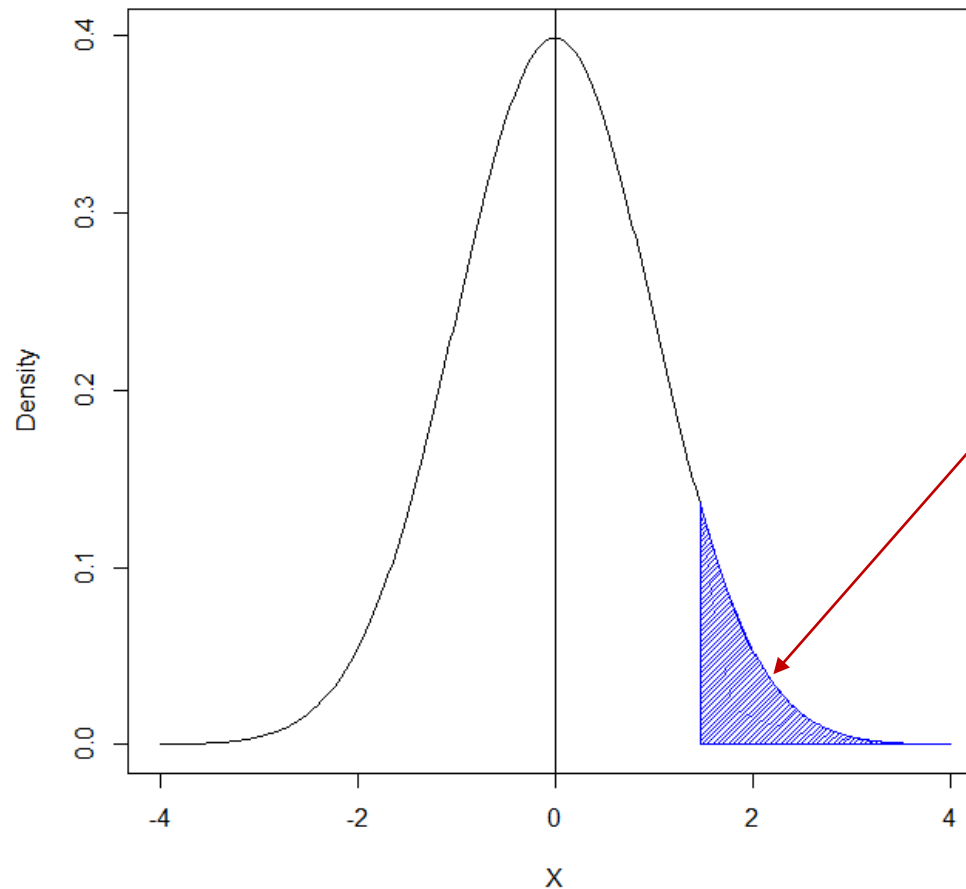
Find $P(X \leq 65)$ and $P(40 \leq X \leq 60)$.

$$P\left(Z \leq \frac{65 - 70}{10}\right) = P(Z \leq -0.5) = 1 - P(Z \leq 0.5) =$$

$$P(40 \leq X \leq 60) = P\left(\frac{40 - 70}{10} \leq Z \leq \frac{60 - 70}{10}\right) = P(-3 \leq Z \leq -1) =$$

z_α - notation

z_α - value on the x-axis for which the area under the curve lies to the right of z_α



Shaded area:

$$P(Z \geq z_\alpha) = \alpha$$

It follows that:

$$P(Z < z_\alpha) = 1 - P(Z \geq z_\alpha) = 1 - \alpha$$

z_α values can be found in normal tables
(see Rosner, page 818) or can be computed
using statistical software

Percentiles: Examples

Let $Z \sim N(0,1)$. Find z_α for the following situations:

1. $P(Z < z_\alpha) = 0.9278$

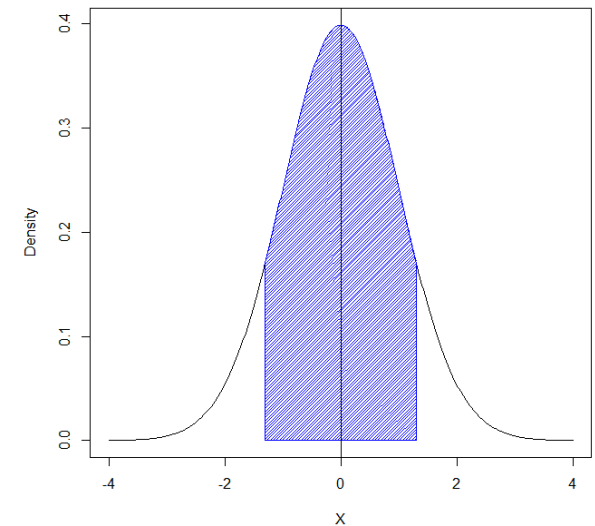
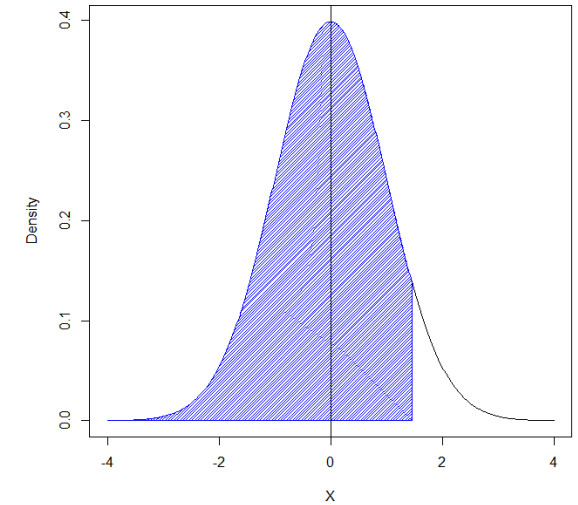
From the normal tables (column A) $\Rightarrow z_\alpha = 1.46$

2. $P(-z_\alpha < Z < z_\alpha) = 0.8132$

$$P(-z_\alpha < Z < z_\alpha) = 2 \cdot P(0 < Z < z_\alpha) = 0.8132$$

$$P(0 < Z < z_\alpha) = \frac{0.8132}{2} = 0.4066$$

From the normal tables (column C) $\Rightarrow z_\alpha = 1.32$



Normal Approximation to Binomial

Let X be a binomial r. v. based on n trials and probability of success p . If the binomial distribution is not too skewed, X may be approximated by a normal distribution under the following two conditions:

$$np \geq 10$$

$$n(1 - p) \geq 10$$

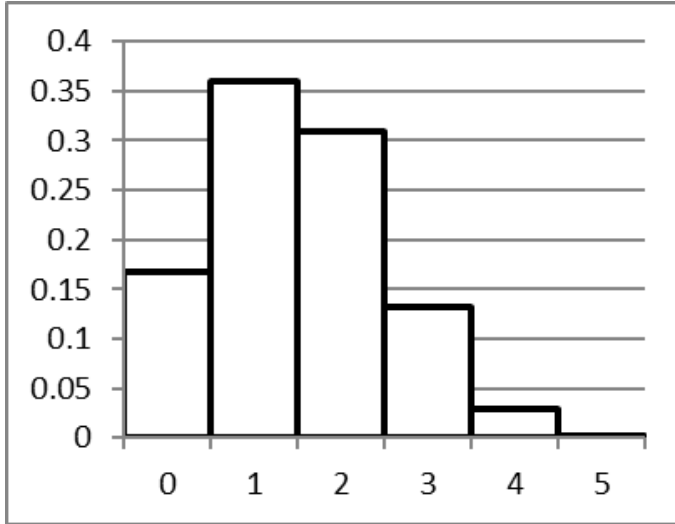
Notes:

- The normal approximation is easier to apply, especially if n is quite large
- Because we are approximating discrete probabilities using a continuous distribution, a continuity correction needs to be applied:

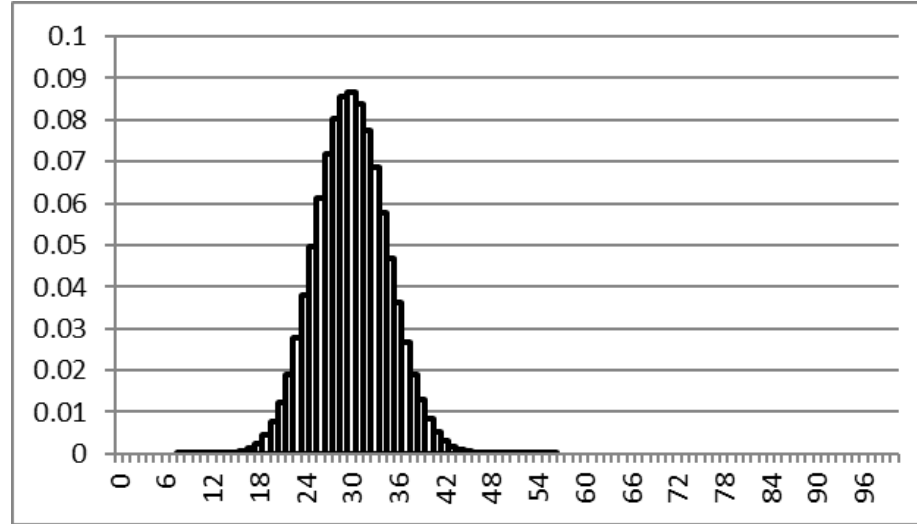
$$P\left(a - \frac{1}{2} \leq X \leq b + \frac{1}{2}\right)$$

Normal Approximation to Binomial: Examples

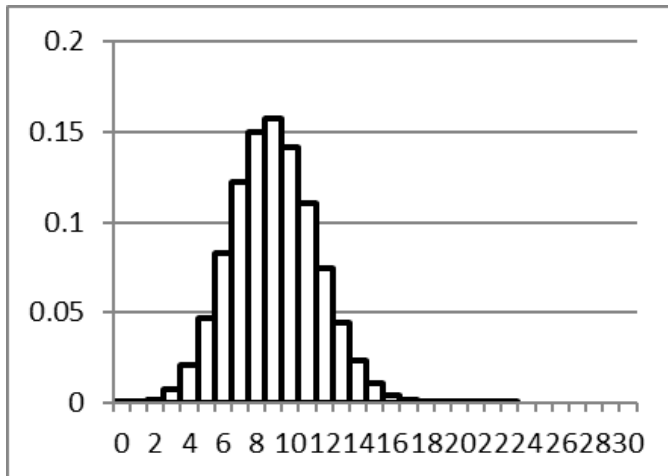
$P=0.30, n=5$



$P=0.30, n=100$



$P=0.30, n=50$



$P=0.03, n=100$



Normal Approximation to Poisson

A Poisson distribution with parameter λ may be approximated by a normal distribution with mean and variance both equal to λ (approximation recommended for $\lambda \geq 10$).

$P(X = x)$ is approximated by:

1. The area under $N(\lambda, \lambda)$ from $x - \frac{1}{2}$ to $x + \frac{1}{2}$, for $x > 0$
2. The area to the left of $\frac{1}{2}$ for $x = 0$.

All normal approximations are based on Central Limit Theorem (CLT). More details about CLT will be provided in Recitation 2.

Readings

Rosner, *Fundamentals of Biostatistics*, Chapter 5

- Sections: 5.2 – 5.5, 5.7 - 5.8
- More details on 5.6 and Central Limit Theorem (CTL) in Recitation 2