Chapter 4

Continuous Random Variables and Probability Distributions

Chaoyue Liu

Department of Mathematics and Statistics

DALHOUSIE UNIVERSITY

Outline

- Continuous Random Variables
- PDFs and CDFs of Continuous Random Variables
- Expected Value and Variance of Continuous Random Variables

- Uniform Distribution
- Exponential Distribution

- Normal Distribution, Standardization and Z-table

Introduction to Continuous Random Variables

Chaoyue Liu

Department of Mathematics and Statistics

DALHOUSIE UNIVERSITY

Outline

- Introduction to Continuous Random Variables
- Probability Density Function (pdf)
- Cumulative Distribution Function (cdf)
- Expected Values and Variance

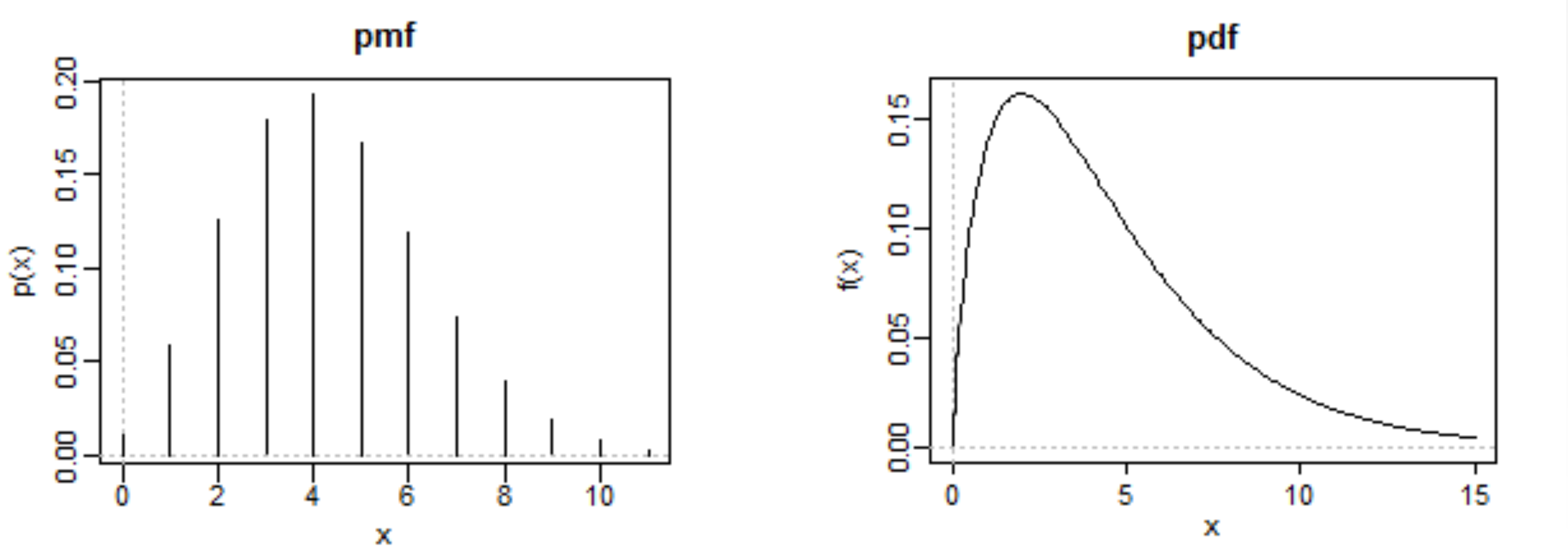
Continuous Random Variables

- Discrete Random Variables: countable values
- Continuous Random Variables: uncountable values
- A random variable X is continuous if:
  - its set of possible values is an entire interval of numbers;
- Examples:
  - The lifetime of a product
  - The waiting time
  - other measured data: length, height, weight etc…

Probability Density Function (pdf)

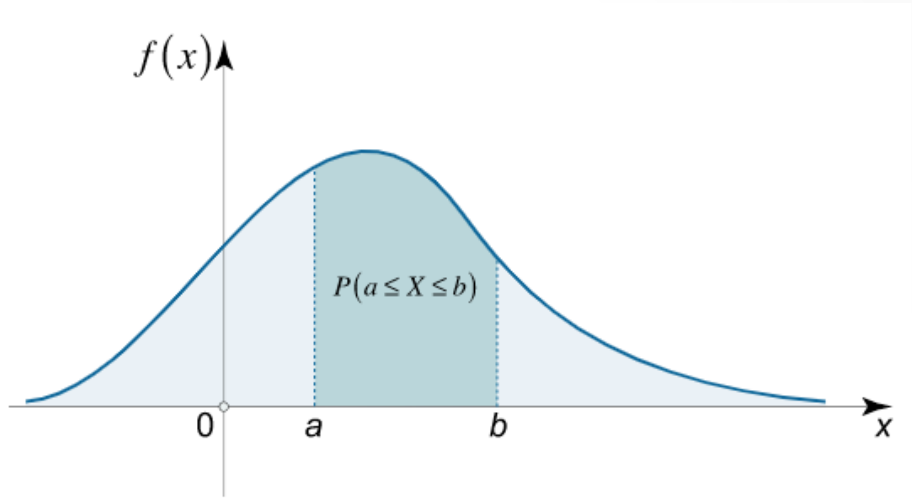
- **Discrete**: probability mass function (pmf)
- **Continuous**: probability density function (pdf)
- Let X be a continuous random variable. The probability density function (pdf) of X is a real valued function $f(x)$ that satisfies
  - $f(x) \geq 0$ for any $x \in \mathbb{R}$
  - For any two real numbers $a \leq b \in \mathbb{R}$, $P(a \leq X \leq b) = \int_a^b f(x)dx$
  - $\int_{-\infty}^{\infty} f(x)dx = 1$

Example:



Probability Density Function (pdf)

- pdf $f(x)$ is not a probability
- The probability for a continuous random variable is given by areas under pdf,



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

- We only talk about the probability of a continuous rv taking the value in an interval, not at a point.
- $P(X = c) = 0$ for any number $c \in \mathbb{R}$.
- Equal sign doesn't matter to continuous rv:

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

DALHOUSIE UNIVERSITY

Cumulative Distribution Function

- Let X have pdf $f(x)$, then the cdf $F(x)$ is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt, \quad \text{for } x \in \mathbb{R}$$

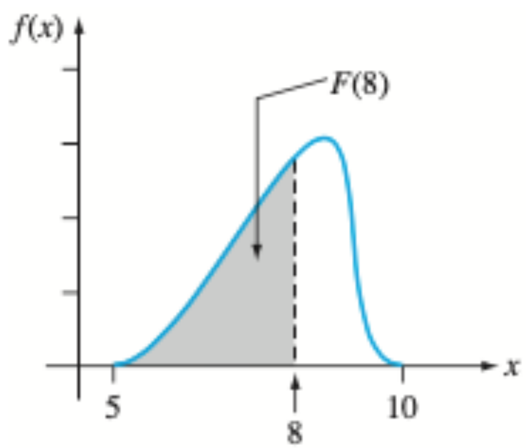
- For $x \in \mathbb{R}$, $F(x)$ is the area under the density curve to the left of x.

- $F(X)$ is non-decreasing

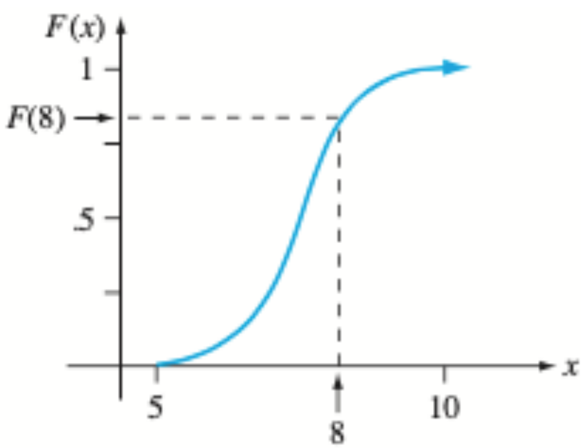
- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

- Example:

pdf



cdf



Relationship between PDF and CDF for Continuous Random Variables

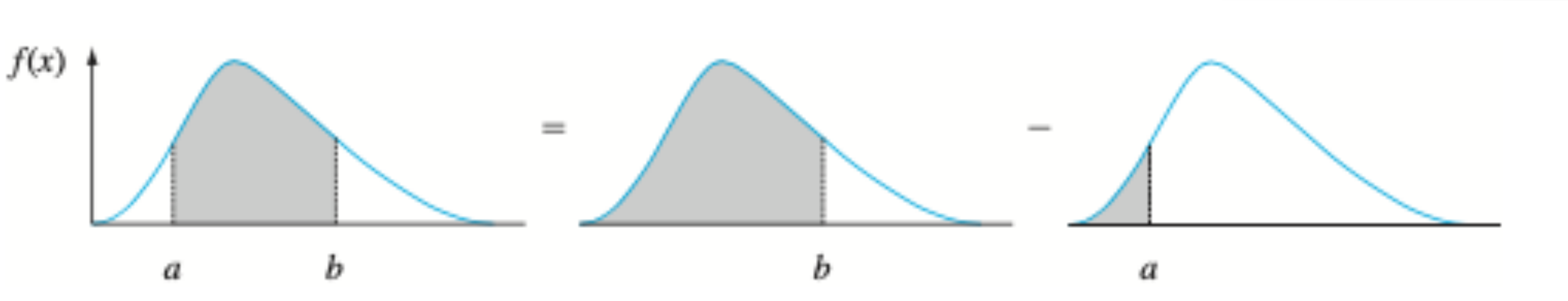
- cdf can be found by integrating the pdf:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

- pdf can be found by differentiating the cdf:

$$f(x) = F'(x) = \frac{d}{dx}F(x)$$

- Compute probability: $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$



Example (pdf -> cdf)

- Given the pdf of X as below,

$$f(x) = \begin{cases} kx^2, & 0 \le x \le 2 \\ 0, & \text{otherwise} \end{cases}$$

1. Find the value of k
2. Find the cdf of X
3. What is the probability that X>1?

(hint: $\int x^n dx = \dfrac{x^{n+1}}{n+1} + c$)

DALHOUSIE UNIVERSITY

Example (cdf -> pdf)

- Let X be a random variable with a cdf:

$$F(x) = \begin{cases} 0, & x \le 0 \\ \frac{x}{4}\left[1 + \ln\left(\frac{4}{x}\right)\right], & 0 < x \le 4 \\ 1, & 4 < x \end{cases}$$

Find:

- (1) $P(X \le 1)$
- (2) $P(1 \le X \le 3)$
- (3) the pdf of X

Percentiles

- A percentile is a value below which a percentage of data falls. e.g. A student's test score is at the 85th percentile of the class means that 85% students scores are lower than that score and 15% are above.
- Let p be a number between 0 and 1. The $(100p)th$ percentile of the distribution of a continuous rv X denoted by $\eta(p)$, is defined by

$$p = P(X \leq \eta(p)) = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(u)du$$

- Special cases:
  - Median ↔ 50th percentile.
  - 1st Quantile ↔ 25th percentile.
  - 3rd Quantile ↔ 75th percentile.

Example

- Let X be a continuous rv with the following pdf:

$$f(x) = 2x,\ 0 \le x \le 1$$

\- Find

- (1) the $100p^{th}$percentile of X
- (2) the median of X

Expected Value and Variance of Continuous Random Variables

- The expected value or mean of a continuous rv X with pdf $f(x)$ is

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

- Let $h(x)$ be a function of X, then

$$\mu_{h(X)} = E(h(X)) = \int_{-\infty}^{\infty} h(u) f(u) du$$

- The variance of X is

$$V(X) = \sigma_X^2 = \sigma^2 = E\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

- The standard deviation of X is $\sigma = \sqrt{\sigma^2}$
- Short-cut formula: $\sigma^2 = E[(X-\mu)^2] = E\left(X^2\right) - \mu^2$
- Linear function of X:
  - Mean: $E[aX+b] = aE[X]+b$ Variance: $V(aX+b) = a^2 V(X)$

Example

- Let X be a random variable with the following pdf
- $$f(x) = \begin{cases} \frac{1}{2} & x \in [0,1] \cup [2,3] \\ 0 & \text{otherwise} \end{cases}$$
  - (1) Check that $f(x)$ is a legitimate pdf
  - (2) Compute $E(X)$ and $V(X)$
  - (3) Compute $P(0.5 \leq X \leq 1.5)$

Discrete vs Continuous

Discrete Random Variables	Continuous Random Variables
pmf : $P(x) = P\{X = x\}$	pdf : $f(x) = F'(x)$
$P\{X \in A\} = \sum_{x \in A} p(x)$	$\boldsymbol{P}\{X \in A\} = \int_A f(x)dx$
$F(x) = \boldsymbol{P}\{X \leq x\} = \sum_{y \leq x} p(y)$	$F(x) = \boldsymbol{P}\{X \leq x\} = \int_{-\infty}^{x} f(y)dy$
$\sum_x p(x) = 1$	$\int_{-\infty}^{\infty} f(x)dx = 1$
$E(X) = \sum_x xp(x)$	$E(X) = \int_{-\infty}^{\infty} xf(x)dx$
$V(X) = \sum_x (x-\mu)^2 p(x)$	$V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$

Uniform Distribution and Exponential Distribution

Chaoyue Liu

Department of Mathematics and Statistics

DALHOUSIE UNIVERSITY

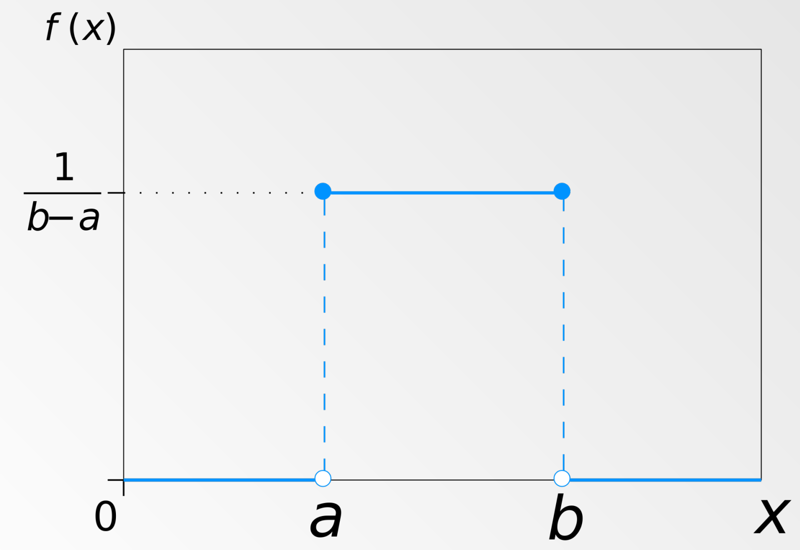
Outline

- Uniform Distribution
- Exponential Distribution

Uniform Distribution

- A continuous random variable X is said to have a uniform distribution on the interval $[a, b]$, denoted by $X \sim uniform(a, b)$, if its pdf is

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



  - Mean: $E(X) = \dfrac{b+a}{2}$
  - Variance: $V(X) = \dfrac{(b-a)^2}{12}$
  - The uniform distribution assigns equal probabilities to intervals of equal lengths
  - For any real numbers $c_1, c_2 \in \mathbb{R}$ such that $a \leq c_1 \leq c_2 \leq b$,

$$P(c_1 \leq X \leq c_2) = \frac{c_2 - c_1}{b - a}$$

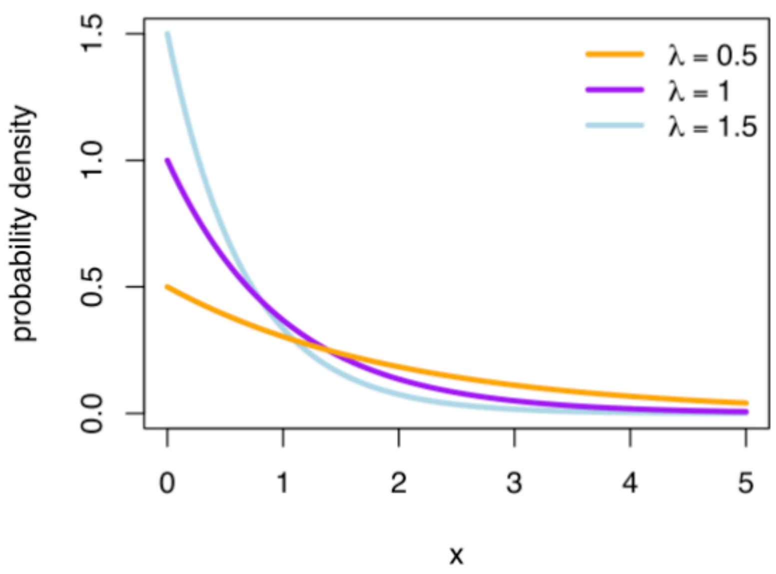
Exercise

- Check that $f(x)$ for Uniform Distribution is a legitimate pdf.

- Derive the cdf of Uniform Distribution.

Exponential Distribution

- A random variable X follows an exponential distribution with (scale) parameter $\lambda > 0$, denoted by $X \sim Exp(\lambda)$, if the pdf of X is



$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- If the number of Independent events occurring follows a Poisson distribution with rate λ, then the distribution of waiting time between two consecutive events follows exponential distribution with same parameter λ.

Properties of Exponential Distribution

- If $X \sim Exp(\lambda)$, then
  - Mean: $E(X) = \frac{1}{\lambda}$
  - Variance: $V(X) = \frac{1}{\lambda^2}$
  - cdf of X: $F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$
  - Memoryless Property, i.e. $P(X > t + s \mid X > s) = P(X > t)$, for any $t, s \geq 0$. i.e. It "forgets" the time already spent.

Example

- Assume that buses arrive at a bus stop with rate α = 4 buses/hour. You get off a bus and wait for the next bus to arrive.
  - (1) What is the probability you wait between 15 to 30 minutes?
  - (2) What is the 40th percentile of your waiting time
  - (3) What is your expected waiting time?

Normal Distribution, Standardization and Z-table

Chaoyue Liu

Department of Mathematics and Statistics

DALHOUSIE UNIVERSITY

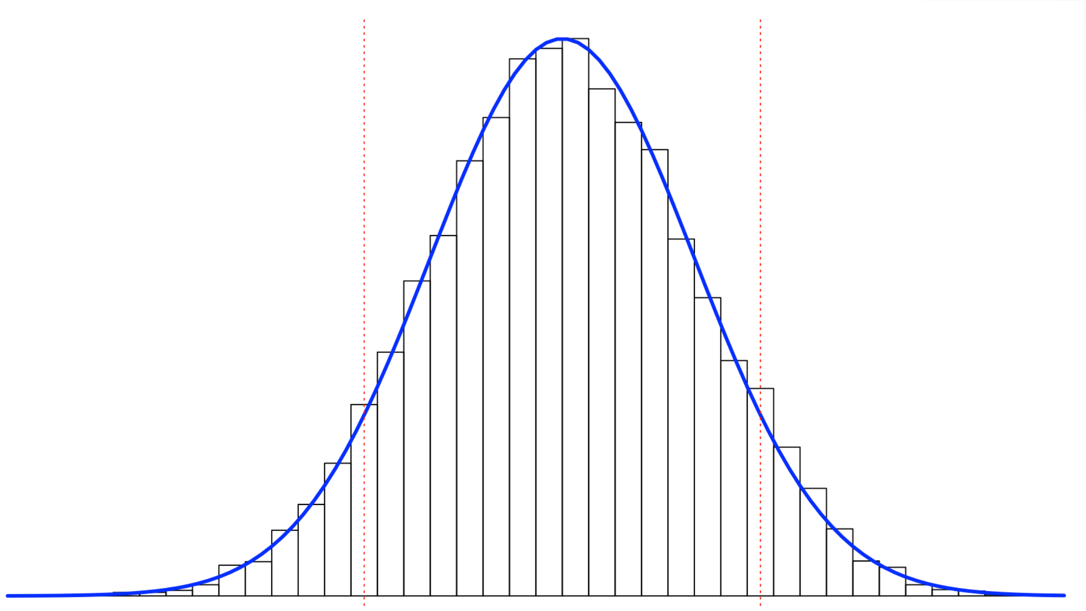
Outline

- Normal Distribution
- Standardization
- Z-table

Normal Distribution

- **Normal distribution**, also known as Gaussian distribution, is very important in statistics, because it occurs naturally in many situations. e.g. height of the population, IQ, measurement error…

- Normal distribution has a symmetric bell-shaped curve (most of values tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur)



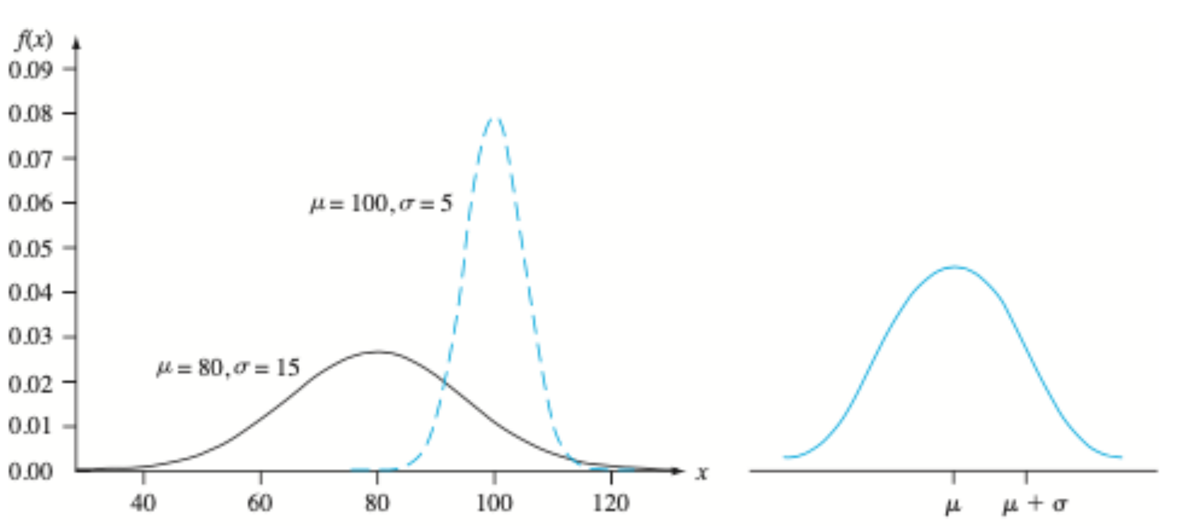
Normal Distribution

- X follows a normal distribution with mean μ and variance σ^2 denoted by $X \sim N(\mu, \sigma^2)$, if the pdf of X is

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

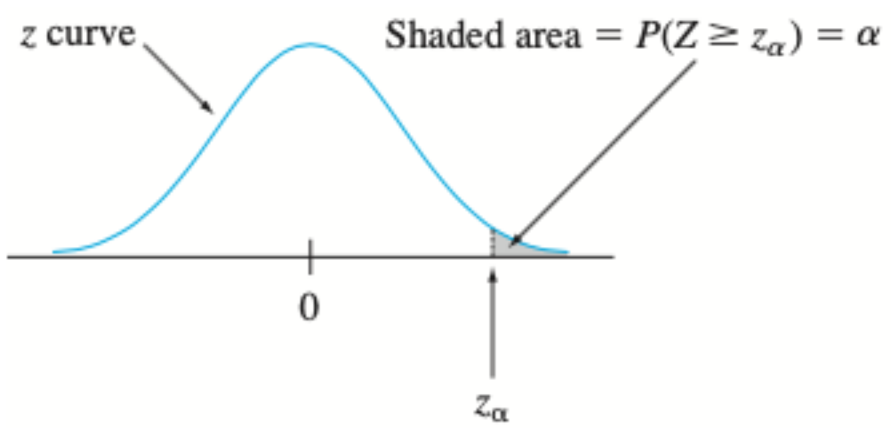
where μ is a real number and σ^2 is a positive number

- $N(\mu, \sigma^2)$ is symmetric at μ
- Mean: $E(X) = \mu$, is a location parameter.
- Variance: $V(X) = \sigma^2$, is a shape parameter. The larger the σ, the more the spread.



Standard Normal Distribution

- The normal distribution with $\mu = 0$ and $\sigma = 1$ is called a standard normal distribution.
- If $X \sim N(\mu, \sigma^2)$, then the random variable $Z = \dfrac{X - \mu}{\sigma}$ will follow the standard normal distribution $Z \sim N(0,1)$. Z is usually called Z value or Z score.
- The cdf of $Z \sim N(0,1)$, usually denoted by $\Phi(z) = P(Z \leq z)$.
- Critical value z_α , is a value such that $P(Z \geq z_\alpha) = \alpha$. z_α is also the $100(1 - \alpha)th$ percentile



e.g.

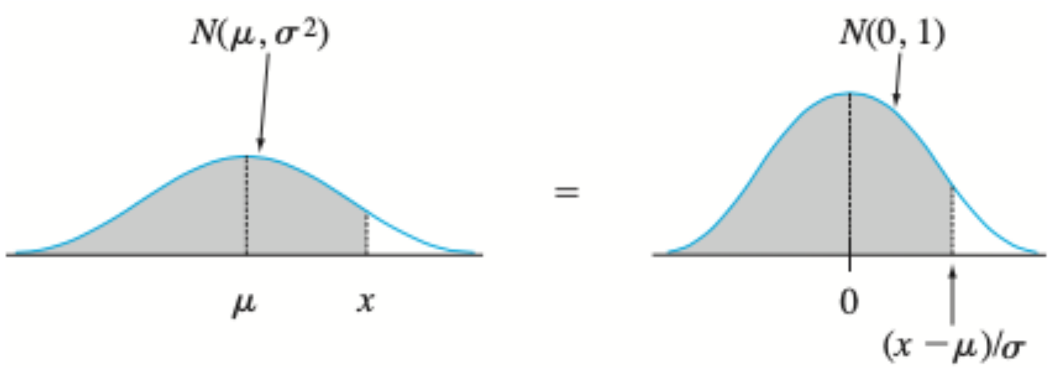
$P(Z \geq z_{0.05}) = 0.05 \Longleftrightarrow P(Z \leq z_{0.05}) = 1 - 0.05 = 0.95$

so $Z_{0.05}$ is also the 95th percentile of Z

Standardization

- Standardization:

If $X \sim N(\mu, \sigma^2)$, then the random variable $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$, and the probability of X can be computed as



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

- compute the cdf of normal distribution is hard in general
- any normal distribution can be converted to standard normal distribution
- one standard normal table (Z table) is only needed to find probabilities for any normal random variables.

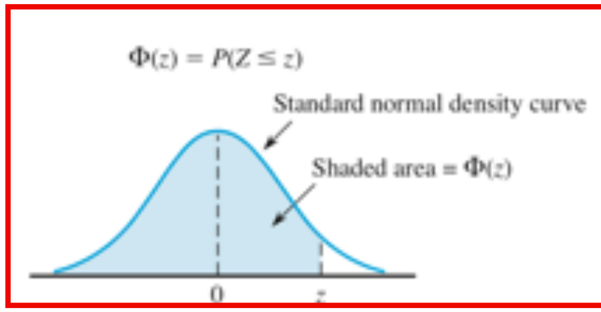
Z table

- Z tables are composed as follows:
  1. The row label contains the integer part and the first decimal place of Z.
  2. The column lable contains the second decimal place of Z
  3. The values within the table are the probabilities

Example:

$$P(Z \le \underbrace{-2.6}_{1}\underbrace{5}_{2}) = \underbrace{0.004}_{3}$$

Table A.3 Standard Normal Curve Areas



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0038

Examples

- given z value, find probability

z	.00	.01	.02	.03	.04
−3.4	.0003	.0003	.0003	.0003	.0003
−3.3	.0005	.0005	.0005	.0004	.0004
−3.2	.0007	.0007	.0006	.0006	.0006
−3.1	.0010	.0009	.0009	.0009	.0008
−3.0	.0013	.0013	.0013	.0012	.0012
−2.9	.0019	.0018	.0017	.0017	.0016
−2.8	.0026	.0025	.0024	.0023	.0023
−2.7	.0035	.0034	.0033	.0032	.0031
−2.6	.0047	.0045	.0044	.0043	.0041
−2.5	.0062	.0060	.0059	.0057	.0055

e.g. what is the probability $P(Z \leq -2.52)$?

- given probability, find z value

z	.00	.01	.02	.03	.04	.05
0.0	.5000	.5040	.5080	.5120	.5160	.5199
0.1	.5398	.5438	.5478	.5517	.5557	.5596
0.2	.5793	.5832	.5871	.5910	.5948	.5987
0.3	.6179	.6217	.6255	.6293	.6331	.6368
0.4	.6554	.6591	.6628	.6664	.6700	.6736
0.5	.6915	.6950	.6985	.7019	.7054	.7088
0.6	.7257	.7291	.7324	.7357	.7389	.7422
0.7	.7580	.7611	.7642	.7673	.7704	.7734
0.8	.7881	.7910	.7939	.7967	.7995	.8023
0.9	.8159	.8186	.8212	.8238	.8264	.8289

e.g. what is the 80th percentile of Z ?

$\Longleftrightarrow \Phi(z) = 0.8$, what is the value of z ?

Examples

- Some useful rules for $Z \sim N(0,1)$
  - $P(Z > a) = 1 - \Phi(a)$
  - $P(|Z| > a) = 2\Phi(-a), a > 0$
  - $P(a < Z < b) = \Phi(b) - \Phi(a)$
  - $P(|Z| < a) = 2\Phi(a) - 1, a > 0$
- note that:
  - $|Z| \geq a \iff Z \leq -a$ or $Z \geq a$
  - $|Z| \leq a \iff -a \leq Z \leq a$
- given α, find the critical value z_α
  - equivalent to be given p, but here $p = 1 - \alpha$
- given $X \sim N(\mu, \sigma^2)$, find the probability of X
  - standardize X first: $Z = \dfrac{X - \mu}{\sigma} \sim N(0,1)$
  - then

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

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Exercise

- Let Z be a standard normal random variable,
  - 1) calculate $P(|Z-1| \leq 1.5)$?
  - 2) If $P(0 \leq Z \leq c) = 0.291$, what is the value of c?
  - 3) determine z_α for $\alpha = 0.2$

Exercise

- Let X be the IQ of a randomly selected person. Assume $X \sim N(100,256)$. What is the probability that a randomly selected person has an IQ
  - above 140?
  - between 92 and 114?
  - Find the median of X.

Normal Approximation to the Binomial Distribution

- If $X \sim Bin(n, p)$ with $np \geq 10$ and $n(1-p) \geq 10$, then X has approximately a normal distribution $N(np, np(1-p))$,

$$P(X \leq x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

  The term 0.5 is added as a correction term for using a continuous distribution to approximate a discrete distribution

- Comparison:
  - Poisson approximation to binomial: n is large and p is small. $n > 50$ and $np < 5$
  - Normal approximation to binomial: not too skewed (enough symmetry). $np \geq 10$ and $n(1-p) \geq 10$.

Exercise

- At a particular college, the pass rate of a course is 72%. If 500 students enroll in a semester, what is the probability that at most 375 students pass?