Theory Notes

Note: Finished Lecture 13

Introduction

- **Probability** is a branch of mathematics concerned with the study of *random phenomenon* (e.g., experiments, models of populations).
- We are primarily interested in probability as it relates to **statistical inference**, the science of drawing inferences about populations based on only a part of the population (i.e., a sample).

Some Definitions

1. **population**: the entire set of objects that we are interested in studying
   e.g., all ISU students
2. **sample**: the subset of the population available for observation
   e.g., STAT 542 students

Note: population and sample are crucial terms in understanding statistics (i.e., STAT 543), but will not occur very often in our discussions of probability theory (i.e., STAT 542).

3. **experiment**: process of obtaining an observed result of a random phenomenon
4. **sample space** S: the set of all possible outcomes of the experiment

- elements $s \in S$ of a sample space are called **sample points** (s)
- a sample space may be
- **discrete**
  (finite or countably infinite, i.e., listable as a finite/infinite sequence)

$$S = \{s_1, s_2, \ldots, s_n\}$$

or

$$S = \{s_1, s_2, s_3, \ldots\}$$

- or **continuous**
  (uncountably infinite, i.e., a continuum of sample points like $S = [0, \infty)$)

5. **event** (e.g., $A, B, \ldots$): subset of the sample space S

- **set**: A is a collection of elements
  (in our case, A is a collection of outcomes)
- **membership**: $x \in A$ or $x \notin A$
  (x is in A or x is not in A)
- **complement**:
  $$A^c = \{x : x \notin A\}$$
  (x such that x is not in A)
- **union**:
  $$A \cup B = \{x : x \in A \text{ or } x \in B\}$$
  (x is in A or B or both)
- **intersection**:
  $$A \cap B = \{x : x \in A \text{ and } x \in B\}$$
- **subset**: $A \subset B$ means that A is contained in B
  (formally, $x \in A \Rightarrow x \in B$)
- **equality**: $A = B$ if $A \subset B$ and $B \subset A$
- **empty set**: $\varnothing$

Algebraic Laws

- **commutativity**:

$$A \cup B = B \cup A$$

$$A \cap B = B \cap A$$

- **associativity**:

$$A \cup (B \cup C) = (A \cup B) \cup C = A \cup B \cup C$$

$$A \cap (B \cap C) = (A \cap B) \cap C = A \cap B \cap C$$

- **distributive law**:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

- **DeMorgan's laws**:

$$(A \cup B)^c = A^c \cap B^c$$

$$(A \cap B)^c = A^c \cup B^c$$

Aside on disjoint and partitions

- events A and B are **disjoint** (mutually exclusive) if

$$A \cap B = \varnothing$$

- For a sequence $A_1, A_2, \ldots$ of events, we say $A_1, A_2, \ldots$ are **pairwise disjoint** if

$$A_i \cap A_j = \varnothing \quad \text{for all } i \neq j$$

- $A_1, A_2, \ldots$ is a **partition** of S if the A_i's are pairwise disjoint and exhaustive, that is,

$$\bigcup_{i=1}^{\infty} A_i = S \quad \text{and} \quad A_i \cap A_j = \varnothing \quad \text{for all } i \neq j$$

Probability Functions

- A **probability function** is a function P defined on a Borel field $\mathcal{B}$ of the sample space S that satisfies:
  1. $P(A) \geq 0$ for all $A \in \mathcal{B}$
  2. $P(S) = 1$
  3. If $A_1, A_2, \ldots \in \mathcal{B}$ are *pairwise disjoint*, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

- Any function satisfying the above is a legitimate probability function.

Theorem 1.2.8.
If P is a probability function and A is any set in $\mathcal{B}$, then:

(a)
$$P(\varnothing) = 0$$

(b)
$$P(A) \leq 1$$

(c)
$$P(A^c) = 1 - P(A)$$

Proof of (c) (parts (a) and (b) follow from (c) and the axioms):

Since

$$S = A \cup A^c,$$

and A and A^c are disjoint, by the axioms of probability,

$$P(S) = P(A \cup A^c) = P(A) + P(A^c).$$

Because $P(S) = 1$, we have

$$1 = P(A) + P(A^c),$$

which implies

$$P(A^c) = 1 - P(A).$$

Theorem 1.2.9.
If P is a probability function and A, B are sets in $\mathcal{B}$, then:

(a)

$$P(B \cap A^c) = P(B) - P(B \cap A)$$

(b)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

(c) If $A \subset B$, then

$$P(A) \leq P(B).$$

Theorem 1.2.11.
If P is a probability function, then

(a) For any partition $C_1, C_2, \ldots \in \mathcal{B}$ (i.e., disjoint C_i's and $\bigcup_{i=1}^{\infty} C_i = S$),

$$P(A) = \sum_{i=1}^{\infty} P(A \cap C_i).$$

(b) For any sets $A_1, A_2, \ldots \in \mathcal{B}$,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i).$$

Principle of Inclusion–Exclusion.
For any sets $A_1, \ldots, A_n$,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k-1} \left(\sum_{1 \leq i_1 < \cdots < i_k \leq n} P(A_{i_1} \cap \cdots \cap A_{i_k}) \right).$$

Equivalently,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1\le i<j\le n} P(A_i \cap A_j) + \cdots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} A_i\right).$$

This generalizes

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

and is proven by induction.

Bonferroni's Inequalities.
For any sets $A_1, \ldots, A_n$ and any $m \in \{1, \ldots, n\}$,

- if m is odd,

$$P\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{k=1}^{m} (-1)^{k-1} \left(\sum_{1\le i_1<\cdots<i_k\le n} P(A_{i_1} \cap \cdots \cap A_{i_k})\right),$$

- if m is even,

$$P\left(\bigcup_{i=1}^{n} A_i\right) \ge \sum_{k=1}^{m} (-1)^{k-1} \left(\sum_{1\le i_1<\cdots<i_k\le n} P(A_{i_1} \cap \cdots \cap A_{i_k})\right).$$

In particular,

$$\sum_{i=1}^{n} P(A_i) - \sum_{1\le i<j\le n} P(A_i \cap A_j) \le P\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n} P(A_i).$$

Combinatorics

Permutations / ordered arrangements II.
When selecting r objects from n objects (without replacement), the number of ordered arrangements possible is

$$n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}.$$

Combinations / unordered selections.
The number of ways to choose r objects from n objects (without replacement), where the ordering doesn't matter, is

$$\binom{n}{r} \equiv \frac{n!}{r!\,(n-r)!}.$$

Summary table: number of ways to select r objects from a group of n

	objects chosen **without replacement**	objects chosen **with replacement**
ordered	$\dfrac{n!}{(n-r)!}$	n^r
unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

Conditional Probability

- **Definition:** If A, B are events in S with $P(B) > 0$, then
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$
- In conditioning, B can be thought of as the **updated sample space**, i.e., not all of S is relevant since we know B has occurred.

$P(\cdot \mid B)$ is a probability function that satisfies the usual axioms and properties.

Axioms:

- $P(A \mid B) \geq 0$ for all events A
- $P(B \mid B) = 1$
(B is the updated sample space)
- If $A_1, A_2, \ldots$ are pairwise disjoint events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i \,\middle|\, B\right) = \sum_{i=1}^{\infty} P(A_i \mid B)$$

Some properties:

$$P(A^c \mid B) = 1 - P(A \mid B)$$

$$P(A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P(A_2 \mid B) - P(A_1 \cap A_2 \mid B)$$

It also follows from our definition of conditional probability that

$$P(A \cap B) = P(B \mid A)\, P(A) = P(A \mid B)\, P(B).$$

More generally, for events $A_1, A_2, \ldots, A_n$,

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

It is possible to reverse the conditioning of A and B to obtain **Bayes' rule**:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}.$$

More generally, if $A_1, A_2, \ldots$ is a partition of the sample space S, then we obtain a general version of Bayes' rule:

$$P(A_i \mid B) = \frac{P(B \mid A_i)\, P(A_i)}{\sum_{j=1}^{\infty} P(B \mid A_j)\, P(A_j)}.$$

Independence

If $P(A \mid B) = P(A)$, then the occurrence of B does not affect the probability of A. It then follows that

$$P(A \cap B) = P(A)P(B) \quad \text{and} \quad P(B \mid A) = P(B).$$

We define two events A and B as **independent** if

$$P(A \cap B) = P(A)P(B).$$

More than two events.
$A_1, \ldots, A_n$ are **independent** if and only if, for any subcollection $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ of distinct indices (with any $2 \leq k \leq n$), it holds that

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j}).$$

- If $A_1, \ldots, A_n$ are independent, then

$$P(A_i \cap A_j) = P(A_i)P(A_j) \quad \text{for any } i \neq j.$$

- However,

$$P(A_i \cap A_j) = P(A_i)P(A_j) \text{ for } i \neq j$$

does **not** imply that $A_1, \ldots, A_n$ are independent.

If $A_1, \ldots, A_n$ are independent, then

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2) \cdots P(A_n).$$

However,

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2) \cdots P(A_n)$$

holding does **not** imply that $A_1, \ldots, A_n$ are independent.

The assumption of independence of events allows the computation of joint occurrences of events through simple calculations.

Random Variables

Definition: A **random variable** (r.v.) X is a function defined on a sample space S that associates a real number with each outcome in S.

That is, for each $s \in S$, we have

$$X(s) \in \mathbb{R}.$$

In function notation,

$$X : S \to \mathbb{R}.$$

We usually suppress the dependence of X on $s \in S$ and write

$$X = X(s).$$

We have $P(A)$ defined on events $A \subset S$, which can be used to assign probabilities for events concerning a random variable X on $\mathbb{R}$ ($X : S \to \mathbb{R}$).

Define $P_X(\cdot)$ for events $B \subset \mathbb{R}$ as follows:

$$P_X(B) = P_X(X \in B) = P\big(\{s \in S : X(s) \in B\}\big).$$

$P_X(\cdot)$ satisfies the axioms and is therefore a legitimate probability function.

CDF

Definition.
The **cumulative distribution function** (cdf) of a random variable X, denoted by $F(\cdot)$, is defined by

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

Sometimes written with subscript as $F_X(x)$.

A function $F(x)$, $x \in \mathbb{R}$, is a cdf for some random variable if and only if the following hold:

1. $F(x)$ is a nondecreasing function of x.
2. $$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1.$$
3. $F(x)$ is right continuous, i.e.,
$$\lim_{x \downarrow x_0} F(x) = F(x_0) \quad \text{for any } x_0 \in \mathbb{R}.$$

Discrete Random Variables

Definition.
If a cdf F is a step function (with jumps at a countable collection of points $x_i \in \mathbb{R}$), then we say the distribution described by F is **discrete** (with support or range $x_i \in \mathbb{R}$).

If a random variable X has a cdf $F = F_X$ which is a step function, then we say X is a **discrete random variable**.

Besides the cdf, there are other (equivalent) ways to state the probability distribution for a discrete distribution / discrete r.v. X.

1. **Probability mass function (pmf).**
   The pmf of a discrete random variable X is given by
   $$f(x) = P(X = x) \geq 0, \quad \text{for any } x \in \mathbb{R}.$$
2. **Equivalent characterization via the cdf.**
   The pmf of a discrete r.v. X can also be written as
   $$f(x) = P(X \leq x) - P(X < x) = F(x) - \lim_{y \to x^-} F(y).$$

Continuous Random Variables and Probability Density Functions

- If a cdf F is such that there exists a nonnegative function f satisfying

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

  then the distribution described by F is said to be (absolutely) **continuous** with **probability density function (pdf)** f.

A random variable X with an (absolutely) continuous cdf F, or a pdf f, is said to be a **continuous random variable**.

- If F is (absolutely) continuous, then its derivative at $x \in \mathbb{R}$ is its pdf $f(x)$:

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

- If X is a continuous random variable, then

$$P(X = x) = 0 \quad \text{for any } x \in \mathbb{R}.$$

For $a < b$,

$$P(a < X < b) = P(a \le X \le b) = P(a \le X < b) = P(a < X \le b) = F(b) - F(a) = \int_a^b f(t)\,dt.$$

Properties of Probability Density or Mass Functions

A function $f(x)$ is a pdf (or pmf) for some random variable if and only if

1. $f(x) \ge 0$ for any $x \in \mathbb{R}$
2. $$\int_{-\infty}^{\infty} f(x)\,dx = 1 \quad (\text{or } \sum_x f(x) = 1)$$

Any nonnegative function having a finite integral (or sum) can be turned into a pdf (or pmf) f by dividing by its integral (or sum).

We will write $X \sim f_X(x)$ (or $X \sim F_X(x)$) to denote that X has a distribution given by f (or F).

Computing Probabilities Using a pmf or pdf

To find general probabilities using a pmf or pdf, note that for $A \subset \mathbb{R}$,

Discrete case (using pmf):

$$P(X \in A) = \sum_{x \in A} f_X(x) = \sum_{x \in A,\ f_X(x) > 0} f_X(x)$$

Continuous case (using pdf):

$$P(X \in A) = \int_A f_X(x)\,dx$$

Relating the CDF to the PMF / PDF

Discrete random variable case

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

$$P(a \leq X \leq b) = F(b) - F(a) + f(a)$$

$$P(a \leq X < b) = F(b) - F(a) + f(a) - f(b)$$

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

Continuous random variable case

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

$$P(a \leq X \leq b) = F(b) - F(a)$$

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

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

Equivalently,

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

Functions of a Random Variable

Introduction

- Consider a random variable $X \sim F_X(\cdot)$ and a function
$$g : \mathbb{R} \to \mathbb{R}.$$
(Here, X is a random variable and g may be *any* function.)
- Then
$$Y = g(X)$$
is also a random variable, having its own cdf $F_Y(\cdot)$.

Since Y is a function of X, we can describe the probabilistic behavior of Y in terms of that of X.

- Formally, there is also an inverse mapping g^{-1} defined by
$$g^{-1}(A) = \{x \in \mathbb{R} : g(x) \in A\}, \qquad \text{for any } A \subset \mathbb{R}.$$

Distribution of a Function of a Random Variable

- The distribution of $Y = g(X)$ is completely determined by the distribution of X and the function g.

For any set $A \subset \mathbb{R}$,

$$P_Y(Y \in A) = P_X(g(X) \in A) = P_X(X \in g^{-1}(A)).$$

That is, the distribution of Y depends on the cdf (or pdf/pmf) F_X of X together with the function g.

Support (Range) Under Transformations

- If X has pdf/pmf $f_X(x)$, then the **range (support)** of X is

$$\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}.$$

- If $Y = g(X)$ has pdf/pmf $f_Y(y)$, then the **range (support)** of Y is

$$\mathcal{Y} = \{y \in \mathbb{R} : f_Y(y) > 0\} = \{g(x) : x \in \mathcal{X}\}.$$

Discrete Case

Result.
If X is a discrete random variable with pmf $f_X(x)$ (i.e., X has range

$$\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\},$$

which is either finite or countably infinite), then

$$Y = g(X)$$

is also a discrete random variable with pmf

$$f_Y(y) = P(Y = y) = \begin{cases} \displaystyle\sum_{x \in g^{-1}(\{y\})} f_X(x) = \sum_{x \in \mathcal{X} : g(x) = y} f_X(x), & y \in \mathcal{Y}, \\ 0, & y \notin \mathcal{Y}, \end{cases}$$

where the range (support) of Y is

$$\mathcal{Y} = \{g(x) : x \in \mathcal{X}\} = \{y \in \mathbb{R} : f_Y(y) > 0\}.$$

Continuous Case

For a continuous random variable X, the random variable

$$Y = g(X)$$

will *typically* (but not always) be continuous.

To determine the distribution of Y, one can use either of the following two approaches.

CDF Method

Compute the cdf $F_Y(\cdot)$ of Y:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(g(X) \leq y) \\ &= P(\{x \in \mathbb{R} : g(x) \leq y\}) \\ &= \int_{\{x \in \mathbb{R}: g(x) \leq y\}} f_X(x)\, dx. \end{aligned}$$

This is a general approach, but its success depends on being able to evaluate the integral.

PDF (Transformation) Method

Alternatively, one may compute the pdf $f_Y(\cdot)$ directly using a transformation technique.

This method is **only valid** when the function g is **monotone** or **piecewise monotone**.

Key Result

Theorem 2.1.5 (Monotone Transformation)

If X has pdf $f_X(x)$ and

$$Y = g(X),$$

where the function $g(\cdot)$ has either a **strictly positive** or a **strictly negative** derivative on

$$\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\},$$

then the pdf of Y has support

$$\mathcal{Y} = \{g(x) : x \in \mathcal{X}\},$$

and is given by

$$f_Y(y) = f_X\big(g^{-1}(y)\big) \left| \frac{dg^{-1}(y)}{dy} \right| > 0, \quad \text{for } y \in \mathcal{Y},$$

with

$$f_Y(y) = 0, \quad \text{for } y \notin \mathcal{Y}.$$

Note that unless g is **strictly monotone** (or at least there is a way to break up

$$\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$$

into several intervals on each of which g is strictly increasing or strictly decreasing),
X being a continuous random variable does **not** necessarily imply that

$$Y = g(X)$$

will be a continuous random variable.

Probability Integral Transform (PIT)

This is a famous (and for some purposes very useful) transformation connected with continuous cdfs.

If F is a continuous cdf, then

$$F(x) = \int_{-\infty}^{x} f(t)\,dt, \quad t \in \mathbb{R}.$$

If X has a continuous cdf $F(\cdot)$, then the random variable

$$Y = F(X)$$

is uniformly distributed on $(0, 1)$.

That is, Y has pdf

$$f_Y(y) = \begin{cases} 1, & 0 < y < 1, \\ 0, & \text{otherwise}, \end{cases}$$

and cdf

$$F_Y(y) = \begin{cases} 0, & y \le 0, \\ y, & 0 \le y \le 1, \\ 1, & y \ge 1. \end{cases}$$

Expected Value of a Function of a Random Variable

Definition.
The expected value (or mean) of a random variable $g(X)$, denoted by $\mathrm{E}g(X)$, $\mathrm{E}[g(X)]$, or $\mathrm{E}(g(X))$, is defined as follows.

Discrete case:

$$\mathrm{E}g(X) = \sum_x g(x)\, f_X(x).$$

Continuous case:

$$\mathrm{E}g(X) = \int_{-\infty}^{\infty} g(x)\, f_X(x)\,dx.$$

Existence of the Expectation

The expectation $\mathrm{E}g(X)$ is defined **provided that**

Discrete case:

$$\sum_x |g(x)|\, f_X(x) < \infty,$$

Continuous case:

$$\int_{-\infty}^{\infty} |g(x)|\, f_X(x)\,dx < \infty.$$

(That is, we require $\mathrm{E}[g(X)]$ to be a real, finite number.)

Nonexistence of the Expectation

We say that the expected value (or mean) $\mathrm{E}g(X)$ **does not exist** if

Discrete case:

$$\sum_x |g(x)|\, f_X(x) = \infty,$$

Continuous case:

$$\int_{-\infty}^{\infty} |g(x)|\, f_X(x)\,dx = \infty.$$

Theorem 2.2.5 (Properties of Expectation)

Theorem.
Suppose X is a random variable such that

$$\mathrm{E}|g_1(X)| < \infty \quad \text{and} \quad \mathrm{E}|g_2(X)| < \infty,$$

and let $a, b, c \in \mathbb{R}$ be fixed constants. Then:

1. $$\mathrm{E}[ag_1(X) + b] = a\,\mathrm{E}g_1(X) + b.$$

2. $$\mathrm{E}[ag_1(X) + bg_2(X) + c] = a\,\mathrm{E}g_1(X) + b\,\mathrm{E}g_2(X) + c.$$

3. If $g_1(x) \geq a$ for all x, then
$$\mathrm{E}g_1(X) \geq a.$$

4. If $g_1(x) \leq b$ for all x, then
$$\mathrm{E}g_1(X) \leq b.$$

5. If $g_1(x) \geq g_2(x)$ for all x, then
$$\mathrm{E}g_1(X) \geq \mathrm{E}g_2(X).$$

Invariance of Expectation Under Transformation

Expectations are invariant under transformation.

If

$$Y = g(X),$$

then

$$\mathrm{E}Y = \sum_y y\, f_Y(y) = \sum_y y\, P(Y = y) = \sum_x g(x)\, f_X(x) = \mathrm{E}g(X)$$

in the discrete case.

(In the continuous case, replace sums with integrals.)

That is,

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx = \mathrm{E}g(X).$$

Variance

An important instance of the $\mathrm{E}g(X)$ notation arises when

$$g(X) = (X - \mathrm{E}X)^2.$$

Definition.
The **variance** of a random variable X, denoted $\mathrm{Var}(X)$ or σ_X^2, is

$$\mathrm{Var}(X) = \sigma_X^2 = \mathrm{E}[X - \mathrm{E}X]^2 = \mathrm{E}\big[(X - \mathrm{E}X)^2\big],$$

the expected squared distance between X and its mean $\mathrm{E}X$.

Two Important Variance Facts

1. For any real numbers a, b,
$$\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X).$$
2. $$\mathrm{Var}(X) = \mathrm{E}X^2 - (\mathrm{E}X)^2.$$

Other Moments and Distributional Summaries

Moments are an important summary of a distribution.

1. $$\mu = \mu_X = \mathrm{E}X$$
is often called the **mean**.
2. $$\mu_n' = \mathrm{E}X^n$$
is the nth (raw) moment, provided $\mathrm{E}X^n$ exists, i.e.,

Discrete case:
$$\sum_x |x^n| f_X(x) < \infty,$$

Continuous case:
$$\int_{-\infty}^{\infty} |x^n| f_X(x)\, dx < \infty.$$

3. $$\mu_n = \mathrm{E}\big[(X - \mu)^n\big]$$
is the nth **central moment**, provided $\mathrm{E}X^n$ exists.

(a) $$\mathrm{Var}(X) = \sigma_X^2 = \mathrm{E}\big[(X - \mu)^2\big] = \mu_2$$
is the **variance**.

(b) $$\sigma_X = \sqrt{\mathrm{Var}(X)}$$
is the **standard deviation**.

(c) $$\mu_3$$
is **skewness** (i.e., measures distributional balance around μ).

(d) $$\mu_4$$
is **kurtosis** (i.e., a measure of how long the distributional tails are).

Regarding Moments

1. If $\mathrm{E}X^r$ exists for some $r > 0$, then $\mathrm{E}X^s$ exists for all

$$0 \leq s \leq r.$$

2. If $\mathrm{E}X^r$ does not exist for some $r > 0$, then $\mathrm{E}X^s$ will not exist for any

$$s > r.$$

3. $\mathrm{E}X^2$ exists if and only if $\mathrm{Var}(X)$ exists.

4. For $r > 0$, the existence of $\mathrm{E}X^r$ is a matter of the distribution of X not having **heavy tails**, i.e., X does not assume large values with large probability.

Convergence (more on this later)

- Suppose a random variable X has mgf $M_X(t)$ and suppose $X_1, X_2, \ldots$ are a sequence of random variables, where X_n has mgf $M_{X_n}(t)$ for each $n \geq 1$.

If

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$$

holds for all $t \in (-h, h)$ for some $h > 0$, then the sequence $X_1, X_2, \ldots$ converges (in distribution) to X.

Moment Generating Functions

Other Generating Functions

There is nothing particularly illuminating about mgfs: these are just a technical device that are sometimes useful for proving theorems. In fact, they are not the only transforms for technical purposes or even the most useful (e.g., Fourier transforms / characteristic functions are often more useful in STAT 642).

Recall that the moment generating function (mgf) is defined as

$$M_X(t) = \mathbb{E}[e^{tX}].$$

Cumulant Generating Function

The function $\log M_X(t)$ is called the **cumulant generating function**.

The nth cumulant is given by

$$\frac{d^n}{dt^n} \log M_X(t)\bigg|_{t=0}.$$

- The **first cumulant** is $\mathbb{E}[X]$
- The **second cumulant** is $\mathrm{Var}(X)$

Factorial Moment Generating Function

The function

$$\mathbb{E}[t^X]$$

is called the **factorial moment generating function (fmgf)**.

The nth factorial moment is given by

$$\frac{d^n}{dt^n}\mathbb{E}[t^X]\bigg|_{t=1} = \mathbb{E}\big[X(X-1)\cdots(X-n+1)\big].$$

For a discrete random variable, the fmgf is also called the **probability generating function**.

Interchanging Orders of Limits

We have several times in lecture up to this point interchanged "orders of limits" (i.e., switching the order of derivatives and expectations, or derivatives and summations, or derivatives and integrals).

This interchange is also implicit in using mgfs to compute moments:

$$\begin{aligned}
\frac{d^n}{dt^n}M_X(t)\bigg|_{t=0} &= \frac{d^n}{dt^n}\mathbb{E}[e^{tX}]\bigg|_{t=0} \\
&= \mathbb{E}\bigg(\frac{d^n}{dt^n}e^{tX}\bigg|_{t=0}\bigg) \\
&= \mathbb{E}\Big(X^n e^{tX}\Big|_{t=0}\Big) \\
&= \mathbb{E}[X^n].
\end{aligned}$$

We don't need to worry about the validity of these interchanges here (it's covered in STAT 642, where it makes more sense with the right technical material in hand).

Inequalities

Markov inequality: Suppose X is a random variable and $g(x) \geq 0$. Then, for any $r > 0$,

$$\mathbb{P}(g(X) \geq r) \leq \frac{\mathbb{E}[g(X)]}{r}.$$

Chebyshev inequality: Suppose X is a random variable with mean $\mathbb{E}[X] = \mu$ and variance σ^2. Then, for any $k > 0$,

$$\mathbb{P}(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2},$$

and equivalently,

$$\mathbb{P}(|X-\mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

Convex Functions

Definition: A function $g(x)$ is **convex** if

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$$

for all x, y and all $0 < \lambda < 1$.

Second Derivative Characterization

If g is twice differentiable, then $g(x)$ is convex if

$$g''(x) \geq 0 \quad \text{for all } x.$$

Example:
If $g(x) = x^2$, then $g'(x) = 2x$ and $g''(x) = 2 > 0$, so $g(x)$ is convex.

A function $g(x)$ is **concave** if $-g(x)$ is convex.

Jensen's inequality: Suppose X is a random variable and $g(x)$ is a convex function. Then,

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]).$$

For a **concave** function, the reverse inequality holds:

$$g(\mathbb{E}[X]) \geq \mathbb{E}[g(X)].$$

Common Distributions

Introduction

- Often it is useful to consider structural forms for a pdf f or cdf F, especially for modeling a population.
- In particular, it is possible to specify a family of distributions using a single functional form with one or more free **parameters**.

Discrete Distributions

Bernoulli Distribution

A **Bernoulli trial** is a random variable

$$X \sim \text{Bern}(p),$$

where p is a parameter.

Definition

- There are **two outcomes**, usually labeled 0 and 1 (failure / success).

$$X = \begin{cases} 0, & \text{with probability } 1-p, \\ 1, & \text{with probability } p. \end{cases}$$

- The parameter satisfies $0 < p < 1$ (i.e., $X = 1$ with probability p).

Probability Mass Function

If $X \sim \text{Bern}(p)$, then

$$\mathbb{P}(X = 0) = 1 - p, \qquad \mathbb{P}(X = 1) = p.$$

Equivalently, the pmf can be written as

$$f_X(x) = f_X(x \mid p) = \mathbb{P}(X = x \mid p) = p^x (1-p)^{1-x}, \qquad x = 0, 1.$$

Expectation

The expectation of X is

$$\mathbb{E}[X] = \sum_{x=0}^{1} x\, \mathbb{P}(X = x) = (0)(1-p) + (1)(p) = p.$$

Second Moment

The second moment is

$$\mathbb{E}[X^2] = \sum_{x=0}^{1} x^2\, \mathbb{P}(X = x) = (0)^2(1-p) + (1)^2(p) = p.$$

Variance

The variance of X is

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1-p).$$

Binomial Distribution

The **Binomial distribution** is defined as

$$X \sim \text{Binom}(n, p), \qquad 0 < p < 1,$$

where n is the number of trials and p is the success probability.

Probability Mass Function

The pmf is given by

$$f_X(x) = f_X(x \mid n, p) = \mathbb{P}(X = x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

Motivation

The Binomial distribution arises as the distribution of the **number of successes** in n independent Bernoulli(p) trials.

Let

$$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Bern}(p),$$

where $Y_i = 1$ if the ith trial is a "success" and 0 if it is a "failure." Define

$$X = \sum_{i=1}^{n} Y_i.$$

Then $X \sim \text{Binom}(n, p)$.

Derivation of the pmf

For a given $x = 0, 1, \ldots, n$,

$$\mathbb{P}(X = x) = \mathbb{P}\left(\sum_{i=1}^{n} Y_i = x\right).$$

Since $Y_1, \ldots, Y_n$ are independent,

$$\mathbb{P}(X = x) = \sum_{\substack{(y_1,\ldots,y_n)\in\{0,1\}^n \\ \sum_{i=1}^n y_i = x}} \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n) = \sum_{\substack{(y_1,\ldots,y_n)\in\{0,1\}^n \\ \sum_{i=1}^n y_i = x}} \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i).$$

Because each $Y_i \sim \text{Bern}(p)$,

$$\prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) = p^{\sum_i y_i}(1-p)^{n-\sum_i y_i} = p^x(1-p)^{n-x}.$$

Thus,

$$\mathbb{P}(X = x) = p^x(1-p)^{n-x} \times \big(\text{number of ways to choose exactly } x \text{ components of } (y_1, \ldots, y_n) \text{ to be } 1\big).$$

The number of such configurations is $\binom{n}{x}$, giving

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Mean and Variance

- Mean:
$$\mathbb{E}[X] = np \qquad \text{(proof on next slide)}$$
- Variance:
$$\text{Var}(X) = np(1-p).$$

Moment Generating Function

The moment generating function of X is

$$M_X(t) = \mathbb{E}[e^{tX}] = \big(pe^t + (1-p)\big)^n, \qquad t \in \mathbb{R}.$$

This follows from

$$M_X(t) = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x},$$

and the binomial expansion of $(a+b)^n$.

Negative Binomial

Let

$$X \sim \text{Neg-Binom}(r, p), \qquad r \in \mathbb{Z},\ r \geq 1,\ 0 < p < 1.$$

Probability Mass Function

The pmf is given by

$$\mathbb{P}(X = x) = f_X(x) = f_X(x \mid r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad x = r, r+1, \ldots$$

Motivation

The Negative Binomial distribution describes the distribution of the **number of independent Bernoulli**(p) **trials needed to obtain** r **successes**.

Equivalently, consider a sequence of successes (S) and failures (F) of total length x such that the rth success occurs on the xth trial.

Alternative Parameterization

Define

$$Y = X - r,$$

the **number of failures prior to the** r**th success**. Then Y is also commonly referred to as Negative Binomial.

The pmf of Y is

$$\mathbb{P}(Y = y) = f_Y(y \mid r, p) = \binom{y+r-1}{y} p^r (1-p)^y, \qquad y = 0, 1, 2, \ldots$$

(Note that $\binom{y+r-1}{y} = \binom{y+r-1}{r-1}$.)

Caution

Be careful: **both** random variables X and Y (which are different) are often called "negative binomial" in the literature.

Mean and Variance

For Y (failures before the rth success),

$$\mathbb{E}[Y] = \frac{r(1-p)}{p}.$$

Since $X = Y + r$,

$$\mathbb{E}[X] = \mathbb{E}[Y] + r = \frac{r}{p}.$$

The variance is

$$\text{Var}(Y) = \frac{r(1-p)}{p^2} \qquad \text{and} \qquad \text{Var}(X) = \text{Var}(Y).$$

Moment Generating Function

The mgf of Y is

$$M_Y(t) = \mathbb{E}[e^{tY}] = \left[\frac{p}{1-(1-p)e^t}\right]^r, \qquad t < -\log(1-p).$$

Since $X = Y + r$,

$$M_X(t) = \mathbb{E}[e^{t(X)}] = \mathbb{E}[e^{t(Y+r)}] = e^{rt} M_Y(t).$$

Geometric

Let

$$X \sim \text{Geom}(p), \qquad 0 < p < 1.$$

Relationship to the Negative Binomial

The Geometric distribution is a **special case** of the Negative Binomial distribution with $r = 1$.

Motivation

The Geometric distribution describes the distribution of the **number of independent Bernoulli**(p) **trials needed to obtain the first success**.

Probability Mass Function

The pmf is

$$f_X(x) = f_X(x \mid p) = p(1-p)^{x-1}, \qquad x = 1, 2, 3, \ldots$$

Mean and Variance

The mean is

$$\mathbb{E}[X] = \frac{1}{p}.$$

The variance is

$$\mathrm{Var}(X) = \frac{1-p}{p^2}.$$

Moment Generating Function

The mgf of X is

$$M_X(t) = \mathbb{E}[e^{tX}] = \frac{pe^t}{1-(1-p)e^t}, \qquad t < -\log(1-p).$$

Memoryless Property of the Geometric Distribution The Geometric distribution has the famous **memoryless** property: for any integer $x_0 \geq 0$,

$$\mathbb{P}(X = x_0 + x \mid X > x_0) = \mathbb{P}(X = x).$$

Interpretation

The conditional distribution of the remaining waiting number of trials until the first success, given that we have already waited x_0 trials, is the same as the original distribution of the number of trials until the first success.

Given that each trial is an independent Bernoulli trial, this makes sense: whether we start counting trials at the beginning, or we start counting trials after x_0 trials without success, the distribution of the remaining number of trials needed until the first success should be the same.

Hypergeometric

Let

$$X \sim \text{Hypergeometric}(N, M, K),$$

where N, M, K are integers.

Probability Mass Function

The pmf is given by

$$\mathbb{P}(X = x) = f_X(x) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, \qquad x = 0, 1, \ldots, K.$$

Motivation

Choose K objects **without replacement** from a total population of size N that contains M "special" objects.

- N = total population size
- M = number of special objects
- $N - M$ = number of non-special objects
- X = number of special objects among the K chosen

Support Conditions

The pmf is nonzero only when

$$0 \le x \le K, \qquad x \le M, \qquad K - x \le N - M.$$

Typically, when $N > 2M$ and $M > K$, only the condition

$$0 \le x \le K$$

matters.

Mean

The mean of a Hypergeometric random variable is

$$\mathbb{E}[X] = K\frac{M}{N}.$$

(Interpretation: sample size $\times$ proportion of special objects.)

Variance

The variance is

$$\mathrm{Var}(X) = \frac{KM(N-M)(N-K)}{N^2(N-1)}.$$

Connection between Hypergeometric and Binomial

Hypergeometric (sampling without replacement)

Choose x special objects in a sample of size K drawn **without replacement** from a population where M objects are "special" and $N - M$ are not.

Binomial (sampling with replacement)

Choose x special objects in a sample of size n, where each selected item has probability p of being a special object.

In particular, if sampling *with replacement* from a population of size N containing M special objects, then

$$X \sim \mathrm{Binomial}\left(n = K,\ p = \frac{M}{N}\right).$$

Result

As the population size N becomes large, the **Hypergeometric distribution tends to the Binomial distribution**.

Poisson

Let

$$X \sim \text{Poisson}(\lambda), \qquad \lambda > 0.$$

Probability Mass Function

The pmf is given by

$$\mathbb{P}(X = x) = f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

Motivation

The Poisson distribution is widely used to model **rare-event count data**, for example: - number of car accidents in a county, - number of system failures in a fixed time period.

Useful Identity

Recall that for any real a,

$$e^a = \sum_{k=0}^{\infty} \frac{a^k}{k!}.$$

This identity ensures that the Poisson pmf satisfies: - $f_X(x) \geq 0$ for all x, - $\sum_x f_X(x) = 1$.

Mean

The mean of a Poisson random variable is

$$\mathbb{E}[X] = \lambda.$$

This follows from

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!}.$$

Variance

The variance of a Poisson random variable is

$$\text{Var}(X) = \lambda.$$

One way to derive this is by noting that

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2,$$

and showing that

$$\mathbb{E}[X(X-1)] = \lambda^2.$$

Moment Generating Function

The moment generating function of X is

$$M_X(t) = \mathbb{E}[e^{tX}] = \exp\big(\lambda(e^t - 1)\big), \qquad t \in \mathbb{R}.$$

Continuous Distributions

Uniform

Let

$$X \sim \text{Uniform}(a, b), \qquad -\infty < a < b < \infty.$$

Probability Density Function

The pdf is given by

$$f_X(x) = \frac{1}{b-a}, \qquad a < x < b.$$

Motivation

The Uniform distribution models **equally likely outcomes** over a finite range (a, b).

- a is the lower endpoint of the range
- b is the upper endpoint

Moments

For $r > 0$,

$$\mathbb{E}[X^r] = \frac{1}{b-a}\int_a^b x^r\,dx = \frac{b^{r+1} - a^{r+1}}{(r+1)(b-a)}.$$

Mean

$$\mathbb{E}[X] = \frac{a+b}{2}.$$

Variance

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{(b-a)^2}{12}.$$

Important Case: Uniform(0, 1)

Let

$$U \sim \text{Uniform}(0, 1).$$

Then:

1. $f_U(u) = 1$, for $0 < u < 1$
2. $\mathbb{E}[U] = \frac{1}{2}$
3. $\text{Var}(U) = \frac{1}{12}$

Probability Integral Transform (PIT)

1. If Y has a **continuous cdf** $F_Y(y)$, then

$$U = F_Y(Y) \sim \text{Uniform}(0, 1).$$

2. Conversely, if $U \sim \text{Uniform}(0, 1)$ and $F_Y(y)$ is a continuous cdf, then

$$Y = F_Y^{-1}(U)$$

has distribution F_Y.

(This is useful for simulating random variables.)

Gamma

Let

$$X \sim \text{Gamma}(\alpha, \beta), \qquad \alpha > 0,\ \beta > 0.$$

Probability Density Function

The pdf is given by

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \qquad 0 < x < \infty.$$

Motivation

The Gamma distribution is a **flexible family for positive-valued random variables**.

Parameters

- $\alpha > 0$ is the **shape parameter**
  - If $\alpha < 1$, the density is unbounded near $x = 0$
  - If $\alpha > 1$, the density is zero at $x = 0$
- $\beta > 0$ is the **scale parameter**

If $X \sim \text{Gamma}(\alpha, \beta)$, then

$$Z = \frac{X}{\beta} \sim \text{Gamma}(\alpha, 1).$$

The Gamma Function

The Gamma function is defined as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx, \qquad \alpha > 0,$$

which ensures that $f_X(x)$ integrates to 1.

Properties of the Gamma Function

1. $\Gamma(1+\alpha) = \alpha\Gamma(\alpha)$, for $\alpha > 0$
2. $\Gamma(\alpha) = (\alpha-1)!$, for integers $\alpha \geq 1$
3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

Moments

For $r > 0$,

$$\mathbb{E}[X^r] = \beta^r \frac{\Gamma(\alpha+r)}{\Gamma(\alpha)}.$$

Mean

$$\mathbb{E}[X] = \alpha\beta.$$

Variance

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \alpha\beta^2.$$

Moment Generating Function

The mgf of X is

$$M_X(t) = \mathbb{E}[e^{tX}] = (1-\beta t)^{-\alpha}, \qquad t < \frac{1}{\beta}.$$

Relationship Between Gamma and Poisson (Integer Shape) For integer α,

$$F_X(x \mid \alpha, \beta) = \mathbb{P}(X \leq x) = \mathbb{P}(Y \geq \alpha), \qquad \text{where } Y \sim \text{Poisson}(x/\beta).$$

Continuous Distributions: Gamma — Special Cases

Chi-squared Distribution If $p > 0$ is an integer, then

$$\chi^2_p \sim \text{Gamma}\Big(\alpha = \frac{p}{2},\ \beta = 2\Big),$$

where p is called the **degrees of freedom** parameter.

Exponential Distribution The Exponential distribution is a special case of the Gamma distribution:

$$\text{Exp}(\beta) = \text{Gamma}(\alpha = 1, \beta).$$

The pdf is

$$f_X(x) = \frac{1}{\beta} e^{-x/\beta}, \qquad 0 < x < \infty.$$

The Exponential distribution is commonly used to model **failure times** and has the **memoryless property**:

$$\mathbb{P}(X > s+t \mid X > t) = \mathbb{P}(X > s).$$

Weibull Distribution If

$$X \sim \text{Exp}(\beta)$$

and $\gamma > 0$, then

$$W = X^{1/\gamma} \sim \text{Weibull}(\gamma, \beta).$$

The pdf of W is

$$f_W(w) = \frac{\gamma}{\beta} w^{\gamma-1} e^{-w^\gamma/\beta}, \qquad 0 < w < \infty.$$

The Weibull distribution is a **general failure time distribution**.

Inverse-Gamma Distribution If

$$X \sim \text{Gamma}(\alpha, \beta),$$

then

$$Y = \frac{1}{X}$$

has the **inverse-gamma distribution**, with pdf

$$f_Y(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \left(\frac{1}{y}\right)^{1+\alpha} e^{-1/(\beta y)}, \qquad 0 < y < \infty.$$

Beta

Let

$$X \sim \text{Beta}(\alpha, \beta), \qquad \alpha > 0,\ \beta > 0.$$

Probability Density Function

The pdf is

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1.$$

Motivation

The Beta distribution is a flexible family, often used for modeling **proportions**.

Shape Parameters

Both α and β are **shape parameters**:

1. α determines behavior near $x = 0$

- If $\alpha < 1$, density is unbounded near 0
- If $\alpha > 1$, density is zero at 0

2. β determines behavior near $x = 1$
   - If $\beta < 1$, density is unbounded near 1
   - If $\beta > 1$, density is zero at 1
3. If $\alpha = \beta$, the distribution is symmetric.

Beta Function

The Beta function is defined as

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

Moments

For $r > 0$,

$$\mathbb{E}[X^r] = \frac{B(\alpha + r, \beta)}{B(\alpha, \beta)} = \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + r)}{\Gamma(\alpha + \beta + r)\Gamma(\alpha)}.$$

Mean

$$\mathbb{E}[X] = \frac{\alpha}{\alpha + \beta}.$$

Variance

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \left(\frac{\alpha}{\alpha + \beta}\right)\left(\frac{\beta}{\alpha + \beta}\right)\left(\frac{1}{\alpha + \beta + 1}\right).$$

Related Distribution

If $\alpha = \beta = 1$, then

$$X \sim \mathrm{Uniform}(0, 1).$$

Normal (Gaussian)

Let

$$X \sim N(\mu, \sigma^2), \qquad -\infty < \mu < \infty,\ \sigma > 0.$$

Probability Density Function

The pdf is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right), \qquad -\infty < x < \infty.$$

Motivation

The Normal distribution is the **single most important distribution** in statistics:

- widely used and analytically tractable
- bell-shaped density arises naturally
- Central Limit Theorem (normal distribution is extremely relevant in large samples)

Parameters

- $\mu \in \mathbb{R}$ is the mean: $\mathbb{E}[X] = \mu$
- $\sigma^2 = \text{Var}(X)$ is the variance; σ is the standard deviation

Standard Normal Distribution

Many properties of the Normal distribution are most easily derived using the standard normal distribution $N(0, 1)$.

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

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

2. If $Z \sim N(0, 1)$, then for constants $a, b \in \mathbb{R}$,

$$X = a + bZ \sim N(\mu = a, \sigma^2 = b^2).$$