Linear Statistical Modeling Methods with SAS

Probability Distributions

Xuemao Zhang
East Stroudsburg University

January 24, 2024

Outline

- Basic Concepts of Probability
- Random Variables
- The Normal Distribution, t-distribution, Chi-square distribution and F-distribution
- Probability Distributions Using SAS

Basic Concepts of Probability

Basic Concepts

- An **experiment** is the process by which an observation is made. Some examples: (1) Record an age (2) Toss a die (3) Toss two coins
- A **simple event**, or **sample point** is an event that cannot be decomposed to simpler components. We use letter E with a subscript to denote a simple event.
  - The basic element to which probability is applied.
  - One and only one simple event can occur when the experiment is performed.
- The set of all simple events or sample points of an experiment is called the **sample space**, denoted by S or Ω.
- A **discrete sample space** is one that contains either a finite or a countable number of distinct sample points.
- An (compound) **event** is a collection of one or more simple events.
- An event **occurs** if one of its simple events occurs.

Example. Toss a die. Define the sample space, simple events and two events: $A = \{\text{an odd number }\}$ and $B = \{\text{a number} > 2\}$.

Basic Concepts of Probability

Axioms of probability

Let $P(A)$ be the probability of event A occurs, $A \subseteq S$.

1. Nonnegativity: $P(A) \geq 0$, for every event $A \subseteq S$.
2. Additivity: If A and B are two disjoint or mutually exclusive events, then
$$P(A \cup B) = P(A) + P(B).$$
3. Normalization: $P(S) = 1$.

Remarks about additivity.

- Finite additivity. If $E_1, E_2, \cdots, E_k$ are simple events, then
$$P(\{E_1, \cdots, E_k\}) = P(E_1) + \cdots + P(E_k).$$
Therefore, $P(A)$ is found by adding the probabilities of all simple events contained in A.
- Countable additivity. If $A_1, A_2, A_3, \ldots$ form a sequence of pairwise mutually exclusive events in S (that is, $A_i \cap A_i = \emptyset$ for $i \neq j$), then
$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

Random Variables

Random variable

A random variable is numerical variable (typically represented by uppercase letters such as X, Y, $\ldots$) that has a single numerical value, determined by chance, for each outcome of an experiment.
Mathematically a **random variable** is a function from a sample space S into the real numbers though we usually ignore this fact and consider the values only of the function.

Notation. Random variables will always be denoted with uppercase letters (e.g. X) and the realized values (e.g. x) of the variable (or its range) will be denoted by the corresponding lowercase letters.

Random Variables

Examples. Random variables in some experiments.

Experiment	Random variable
1 coin toss	$X(H) = 1, X(T) = 0$
Toss two dice	$X =$ sum of the numbers
Toss a coin 25 times	$X =$ number of heads in 25 tosses

Discrete Random Variable

A random variable Y is said to be discrete if it can assume only a finite or countably infinite number of distinct values.

Continuous Random Variable

A random variable Y is said to be continuous if it can assume infinitely many values corresponding to the points on a real line interval.

Cumulative distribution function

cumulative distribution function

The cumulative distribution function or cdf of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P(X \leq x) \text{ for all } x.$$

Properties of cdf

The function $F(x)$ is a cdf if and only if the following three conditions hold:

1. $F(-\infty) = \lim_{x\to-\infty} F(x) = 0$ and $F(\infty) = \lim_{x\to\infty} F(x) = 1$.
2. $F(x)$ is a non-decreasing function of x.
3. $F(x)$ is right-continuous; that is, for every number x_0, $\lim_{x\to x_0^+} F(x) = F(x_0)$.

We have more rigorous definitions of a continuous and discrete random variable in terms of CDF.

A random variable X is **continuous** if its cdf $F(x)$ is a continuous function of x.
A random variable X is **discrete** if its cdf $F(x)$ is a step function of x.

Probability density function

probability density function

The probability density function or pdf, $f_X(x)$, of a continuous random variable X is the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \text{ for all } x.$$

Note.

1. If cdf $F(x)$ is known, then pdf $f(x) = \dfrac{dF(x)}{dx} = F^{'}(x)$.
2. A pdf $f(x) \geq 0$.

probability density function

A function $f(x)$ is a pdf of a random variable X if and only if
(a) $f(x) \geq 0$ for all x.
(b) $\int_{-\infty}^{\infty} f(t)dt = 1$.

Expected Values

Expected value

Let Y be a continuous random variable with the probability density function $f(y)$. Then the expected value of Y , denoted by $E(Y)$ or μ, is defined to be

$$\mu = E(Y) = \int_{-\infty}^{\infty} y f(y) dy$$

provided that the integral exists.

Expected Value of a Transformation

If Y is a continuous random variable and g is a function, then

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y) f_Y(y) dy.$$

Remark. In general, $E[g(Y)] \neq g[E(Y)]$.

Variance

$\sigma^2 = Var(Y) = E[(Y - \mu)^2] = \int_{-\infty}^{\infty} (y - \mu)^2 f_Y(y) dy = E(Y^2) - \mu^2$

Expected Values

Properties

Let Y be a continuous random variable with probability density function $f(y)$, mean μ and variance σ^2.

1. $E(c) = c$ for any constant c.
2. $E[cg(Y)] = cE[g(Y)]$ for any function g of Y and constant c.
3. $E\left(\sum_{i=1}^{k} g_i(Y)\right) = \sum_{i=1}^{k} E[g_i(y)]$.
4. $E(c_1 Y + c_2) = c_1 E(Y) + c_2$ for any constants c_1 and c_2.
5. $Var(c) = 0$ for any constant c.
6. $Var(cY + b) = c^2 Var(Y)$ for any constant c and b.

The Normal Distribution

Definition. A random variable Y is said to have a normal probability distribution if and only if, for $\sigma > 0$ and $-\infty < \mu < \infty$, the pdf of Y is

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

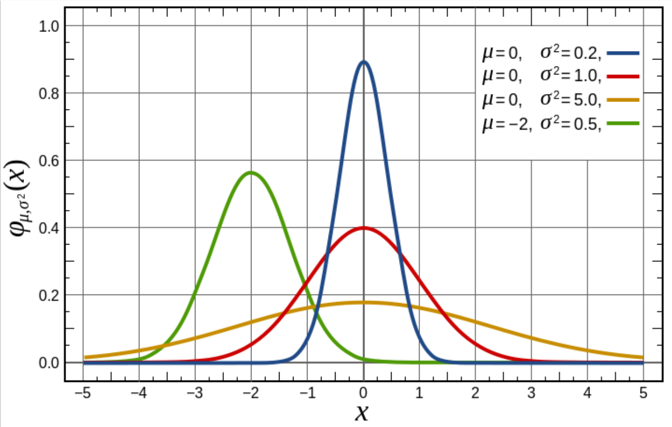
Theorem. If Y is a normally distributed random variable with parameters μ and σ, then

$$\boxed{E(Y) = \mu \text{ and } Var(Y) = \sigma^2.}$$

Theorem. Let $Y \sim N(\mu, \sigma^2)$. Then

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

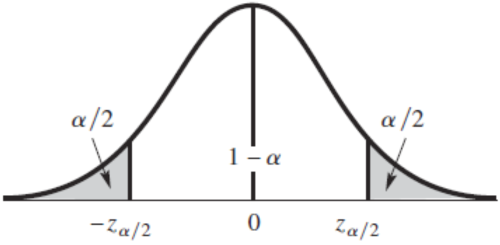
The Normal Distribution



1. Mean $= \mu$; Standard deviation $= \sigma$.
2. Symmetric about $x = \mu$.
3. Total area under the curve is 1.

The Normal Distribution

Definition. The Z **critical value*}** with right tail area α is a point, denoted by Z_α, such that $P(Z \geq Z_\alpha) = \alpha$.



Distribution of *Z*

Some special Z critical values

- $Z_{0.05} = 1.645$.
- $Z_{0.025} = 1.96$.
- $Z_{0.01} = 2.325$.
- $Z_{0.005} = 2.575$.

Student's t-Distribution

DEFINITION. Let Z be a standard normal random variable and let W be a χ^2-distributed variable (to be discussed) with v df. If Z and W are independent, then

$$T = \frac{Z}{\sqrt{W/v}}$$

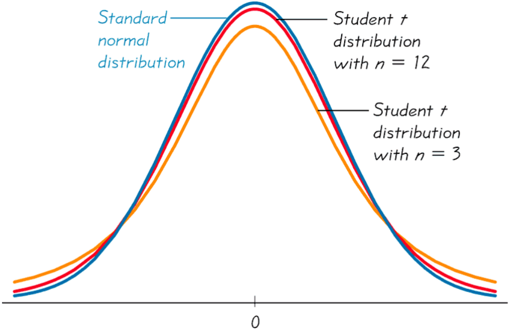
is said to have a t-distribution with v df.

Theorem. Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

has Student's t-distribution with $n - 1$ degrees of freedom.

Student's t-Distribution



1. The density curves of the t-distribution look quite similar to the standard normal curve.
2. The spread of the t-distributions is a bit bigger than that of the standard normal curve.
3. As df gets bigger, the $t(df)$ density curve gets closer to the standard normal density curve.

Chi-square Distribution

THEOREM. Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then $Z_i = (Y_i - \mu)/\sigma$ are independent, standard normal random variables, $i = 1, 2, \ldots, n$, and

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

has a χ^2 distribution with n degrees of freedom (df).

THEOREM. Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Let $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$ be the sample variance. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2}$$

has a χ^2 distribution with $n-1$ degrees of freedom (df). Also, $\overline{Y}$ and S^2 are **independent** random variables.

Chi-square Distribution

THEOREM. Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then $Z_i = (Y_i - \mu)/\sigma$ are independent, standard normal random variables, $i = 1, 2, \ldots, n$, and

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

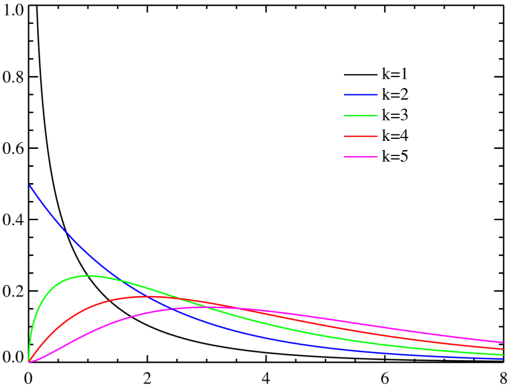
has a χ^2 distribution with n degrees of freedom (df).

THEOREM. Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Let $S^2 = \dfrac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$ be the sample variance. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2}$$

has a χ^2 distribution with $n-1$ degrees of freedom (df). Also, $\overline{Y}$ and S^2 are **independent** random variables.

Chi-square Distribution



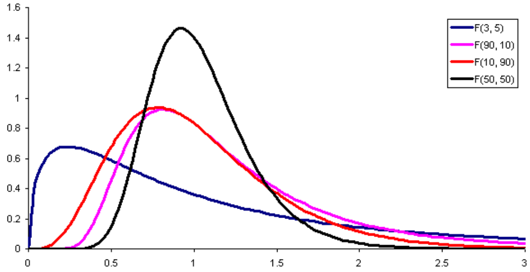
1. The values of chi-square can be zero or positive, but it cannot be negative.
2. The chi-square distribution is not symmetric, unlike the Normal distributions. As the number of degrees of freedom increases, the distribution approaches a Normal distribution and thus becomes more symmetric.

F-Distribution

DEFINITION. Let W_1 and W_2 be independent χ^2-distributed random variables with v_1 and v_2 df, respectively. Then,

$$F = \frac{W_1/v_1}{W_2/v_2}$$

is said to have an F distribution with v_1 numerator degrees of freedom and v_2 denominator degrees of freedom.



1. The F distribution is not symmetric.
2. Values of the F distribution cannot be negative.
3. The exact shape of the F distribution depends on the two different dfs:

CDF calculations

- **Example 1**. X1 $\sim$ binomial (n=12, p=0.67). Find P(X1 $\leq$ 5).

```
data a1;
y = cdf('BINOMIAL', 5, 0.67, 12);
run;

proc print data=a1;
run;
```

CDF calculations

- **Example 2**. X1 ~ binomial (n=12, p=0.67). Find P(X1 = 5).

```
data a2;
y1 = cdf('BINOMIAL', 5, 0.67, 12);
y2 = cdf('BINOMIAL', 4, 0.67, 12);
y = y1 - y2;
run;

proc print data=a2;
run;
```

CDF calculations

- **Example 3**. X2 ~ Poisson (μ=4.35). Find P(X2 $\leq$ 5).

```
data a3;
y = cdf('POISSON', 5,  4.35);
run;

proc print data=a3;
run;
```

CDF calculations

- **Example 4**. Y $\sim$ Normal (μ=0.8, σ=1.15). Find P(Y $\leq$ 2.2) and P(Y $\geq$ 1.8).

```
data a4;
y1 = cdf('NORMAL', 2.2,  0.8, 1.15);
y2 = 1- cdf('NORMAL', 1.8,  0.8, 1.15);
run;

proc print data=a4;
run;
```

Quantile calculations

- Given a cumulative probability (Lower tail) or 1- cumulative probability (Upper tail), we want to see the corresponding value of a random variable. This is the problem of finding the quantitle of a random variable. For example, find a z-score or a score of a non-standard normal random variable.
- We are especially interested in finding quantitles of a continuous random variable.

Quantile calculations

- **Example 1**. Y $\sim$ Normal (μ=0.8, σ=1.15).

1. If P(Y $\leq$ y1) = 0.775, y1=?

```
data b1;
y=quantile('NORMAL',0.775, 0.8, 1.15);
y2= quantile('NORMAL',0.975, 0, 1);
run;

proc print data=b1;
run;
```

Quantile calculations

2. If $P(Y \geq y2) = 0.662$, y2=?

```
data b2;
y=quantile('NORMAL',1-0.662, 0.8, 1.15);
run;

proc print data=b2;
run;
```

Quantile calculations

- **Example 2.** $Y \sim t$ (df=15). If $P(Y \leq y) = 0.975$, y=?

```
data b3;
y=quantile('T',0.975, 15);
run;

proc print data=b3;
run;
```

Random number generation

- We use RAND function in SAS
- **Example**. Generate 100 standard normal random numbers.

```
data normal (keep=x); /* keep the random numbers only */
call streaminit(4321); /*set the seed value using STREAMINIT functio
do i=1 to 100;
x=rand('NORMAL', 0, 1);
output; /* output the random numbers */
end;   /* do loop */
run;

proc print data=normal;
run;
```

License