Basic of Distribution

listenzcc

December 8, 2020

Abstract

A family of *normal* distributions, like Normal, Chi-squared and Student's t-distribution. The Normal Distribution is the *core* conception, all others are derived from it.

- This article begins with Gamma and Beta function. Since they are useful for computing the *moments* the normal distribution family.
- Then the distributions are described one by one.
  - Normal distribution
  - Chi-squared distribution
  - Student's t-distribution
- The Appendix provides the necessary proofs.

Contents

1 Prepare Knowledge

1.1 Gamma and Beta function

An infinity integral is called as *Gamma (Γ) function*

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \, \mathrm{d}x \tag{1}$$

The *Beta (B) function* is a two-factor function, derived from Γ function

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{2}$$

1.2 Important equations

Proposition 1.1. *Some very important equations.*

The value of $\Gamma(\frac{1}{2})$

$$\Gamma(\frac{1}{2}) = \sqrt{\pi} \tag{3}$$

The recursive of $\Gamma(n)$*, the general situation,*

$$\Gamma(1+z) = z\Gamma(z) \tag{4}$$
$$\Gamma(1-z) = -z\Gamma(-z) \tag{5}$$

The integer situation,

$$\Gamma(n) = (n-1)! \quad \forall n \in \mathcal{N}^+ \tag{6}$$

The relationship between Γ *and* e^{-x^2}

$$\Gamma(z) = 2\int_0^\infty x^{2z-1} e^{-x^2} \, \mathrm{d}x \tag{7}$$

The relationship between Γ *and* B *Function*

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1} (1-t)^{\beta-1} \, \mathrm{d}t \tag{8}$$

See Lemma A.1 for proof.

2 Normal Distribution

2.1 Definition

It is hard to say normal distribution is what, since almost every thing follows it.

The Probability Distribution Function *(pdf)* of normal distribution is

$$p(x) = \frac{1}{\sqrt{2\pi}\delta} \exp(-\frac{(x-\mu)^2}{2\delta^2}), -\infty < x < \infty \tag{9}$$

the symbolic notion is $p(x) \sim \mathcal{N}(\mu, \delta^2)$. When $\mu = 0$ and $\delta^2 = 1$, it is called standard normal distribution.

2.2 Mean and Variance

The mean and variance of the normal distribution is

$$Mean \triangleq E(x) = \mu$$
$$Variance \triangleq E(x^2) - E^2(x) = \delta^2$$

it is easy to proof using Proposition 1.1.

3 Chi-squared Distribution

3.1 Definition

If $Y_i \sim \mathcal{N}(0,1)$, then

$$\chi^2 \equiv \sum_{i=1}^{r} Y_i^2 \tag{10}$$

is distributed as Chi-squared χ^2 distribution with r degrees of freedom. The symbolic notion is $p_r(x) \sim \chi^2(r)$.

The pdf of Chi-squared distribution is

$$p_r(x) = \frac{x^{r/2-1}e^{-x/2}}{\Gamma(r/2)2^{r/2}}, 0 < x < \infty \tag{11}$$

The proof can be found in Lemma A.2 and Lemma A.3.

3.2 Relationship with Normal Distribution

The Chi-squared distribution is derived from Normal Distribution. The relationship is not direct, but it is essential to Student's t-distribution, which is low-sample version of Normal Distribution.

3.3 Mean and Variance

The mean and variance of the chi-squared distribution is

$$Mean \triangleq E(x) = r$$
$$Variance \triangleq E(x^2) - E^2(x) = 2r$$

it is easy to proof using Proposition 1.1.

4 Student's t Distribution

4.1 Definition

The probability distribution of a random variable T, of the form

$$T = \frac{\bar{x} - m}{s/\sqrt{N}} \tag{12}$$

where $\bar{x}$ is the sample mean value of all N samples, m is the population mean value and s is the population standard deviation.

Or, in a more formal one

$$T = \frac{X}{\sqrt{Y/r}} \tag{13}$$

where $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_r$.

The pdf of Student's t-distribution is

$$t_r(x) = \frac{\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r}{2})\sqrt{r\pi}}(1 + \frac{x^2}{r})^{-\frac{r+1}{2}}, -\infty < x < \infty \tag{14}$$

it is easy to proof the pdf is a pdf Lemma A.5.

The pdf of Student's t-distribution can be computed using Lemma A.4.

4.2 Relationship with Normal Distribution

It is easy to see that $\lim_{r\to\infty} t_r(x) \sim \mathcal{N}(0,1)$. It demonstrates that when r is large enough, the Student's t-distribution is equalize to Normal Distribution.

4.3 Mean and Variance

The mean and variance of the Student's t-distribution is

$$\begin{aligned} Mean &\triangleq E(x) = 0 \\ Variance &\triangleq E(x^2) - E^2(x) = \frac{r}{r-2} \end{aligned}$$

5 Examples

5.1 Determine the valid sample size

People trend to believe the results of surveys, but are they true? The sample size may be a key to answer the question.

Ground Truth
Let's start with a simple model of 'How do people support a candidate?'. The supporting rate is a random variable that fits

$$R_{support} \sim \mathcal{N}(\mu, \sigma^2)$$

where μ and σ^2 are unavailable.

Population Survey The supporting rate is so important that they want to know it using *LARGE* surveys. It is like asking every body. The number of supporters follows binominal distribution

$$p(n) = (N, n) \cdot r^n \cdot (1 - r)^{N-n}$$

where r refers the ground truth of supporting rate, N is the population and n is the number of supporters.

Then, the sample expectation and sample variance of n is direct

$$E(n) = N \cdot r \tag{15}$$

$$D(n) = N \cdot r \cdot (1 - r) \tag{16}$$

Divide by N, we can get the *unbiased estimation* [1] of r as $\hat{r}$ and its sample variance

$$\hat{r} = \frac{n}{N} \tag{17}$$

$$D(\hat{r}) = \frac{r \cdot (1 - r)}{N} \tag{18}$$

since $E(\hat{r}) = r$.

It turns out that the variance is related to r value. There are things to remember

- The variance is symmetric to 0.5.
- The closer r to 0.5, the larger is the variance.
- The variance decreases when the sample size increases.

Sample Survey
However, in practice, the survey of all population is impossible. Usually, survey in a small group (the number is $M < N$) is available. It will turn out a similar situation

$$\hat{r} = \frac{m}{M} \tag{19}$$

$$D(\hat{r}) = \frac{r \cdot (1 - r)}{M} \tag{20}$$

[1] The unbiased estimation refers the sample mean of the random variable equals to the expectation.

Estimation

Use above analysis, we can say that the estimation of μ value of $R_{support}$ is easy to compute. But the real question is *How we can trust the estimation?*. Especially in the case of the survey is restricted in *part* of the population.

We analysis the question on both side.

- **In ground-truth end**, we have the equation that

$$\frac{\sigma^2}{N} = \frac{s^2}{N-1}$$

  where s^2 refers the sample variance with population of N.

- **In survey end**, we have the variance that

$$s_N^2 = \frac{r(1-r)}{N}$$
$$s_M^2 = \frac{r(1-r)}{M}$$

  where s_N^2 and s_M^2 refer the variance of N and M population.

- **Jointly**, since the variance of M is derived from the intrinsic variance σ^2, and population of N. It meets

$$s_M^2 = s_N^2 + s^2$$

Use the joint equation, the M follows

$$M = N \cdot \frac{\frac{r(1-r)}{\sigma^2}}{(N-1) + \frac{r(1-r)}{\sigma^2}}$$

Under certain error edge e, use the equation between e and σ^2

$$e^2 = z^2 \cdot \sigma^2$$

where z is the *Percentile* of normal distribution according to e.

Thus, we have the relationship between e and M.

$$\hat{M} = N \cdot \frac{z^2 \frac{r(1-r)}{e^2}}{(N-1) + z^2 \frac{r(1-r)}{e^2}}$$

The solution is the minimization of *Sample size* to achieve certain degree of confidence defined by e.

A Appendix

A.1 The relationship between Γ and $B(\alpha, \beta)$

Lemma A.1. *The relationship between Γ and $B(\alpha, \beta)$*

$$\Gamma(m)\Gamma(n) = B(m, n)\Gamma(m + n) \tag{21}$$

Proof. One can write

$$\Gamma(m)\Gamma(n) = \int_0^\infty x^{m-1}e^{-x}dx \int_0^\infty y^{n-1}e^{-y}dy$$

Then rewrite it as a double integral

$$\Gamma(m)\Gamma(n) = \int_0^\infty \int_0^\infty x^{m-1}y^{n-1}e^{-x-y}dxdy$$

Applying the substitution $x = vt$ and $y = v(1 - t)$, we have

$$\Gamma(m)\Gamma(n) = \int_0^1 t^{m-1}(1 - t)^{n-1}dt \int_0^\infty v^{m+n-1}e^{-v}dv$$

Using the definitions of Γ and Beta functions, we have

$$\Gamma(m)\Gamma(n) = B(m, n)\Gamma(m + n)$$

Hence proved.

□

A.2 The pdf of Chi-squared distribution

Lemma A.2. *To get the pdf of a Chi-squared distribution, we have to prove that*

$$p_n(x) \propto x^{n/2-1} \cdot e^{-x/2}$$

in which, $x = \sum_{i=1}^n y_i^2$ and $y_i \sim \mathcal{N}(0, 1)$. Each y_i are independent.

Proof. The joint probability of $\{y_1, y_2, \ldots, y_n\}$ is

$$p_{joint} = exp(\sum_{i=1}^n -y_i^2/2)$$

Thus, the cumulative sum of $p_n(x)$ can be computed using surface integral

$$P_n(r < \sqrt{x}) \propto \int_S p_{joint}ds$$
$$P_n(r < \sqrt{x}) \propto \int_S e^{-r^2/2}ds$$

in which, S refers the volume of a sphere with radius of x.

Transfer the integral into sphere coordinates, we have

$$P_n(r < \sqrt{x}) \propto \int_{r=0}^{\sqrt{x}} e^{-r^2/2}r^{(n-1)}dr$$

Derivate to x, we have

$$\frac{\partial}{\partial x}P_n(r < \sqrt{x}) \propto e^{-r^2/2}r^{(n-1)}x^{-1/2}$$
$$\frac{\partial}{\partial x}P_n(r < \sqrt{x}) \propto x^{n/2-1} \cdot e^{-x/2}$$

because of the Newton's integral rule, the second step is based on the replacement of $r = \sqrt{x}$.

Hence proved.

$\square$

Lemma A.3. *Next, we have to prove that the integral of $p_n(x)$ with $p_n(x) \sim \chi^2(n)$ is*

$$\int_0^\infty p_n(x)dx = \Gamma(n/2) \cdot 2^{r/2}$$

Proof. Use the definition of Γ function

$$\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx$$

Use variable replacement of $z = 2x$, we have

$$\Gamma(n) = 2^{-n}\int_0^\infty z^{n-1}e^{-z/2}dz$$

Then, use substitution of $n = n/2$, we have

$$\Gamma(n/2) \cdot 2^{n/2} = \int_0^\infty z^{n/2-1}e^{-z/2}dz$$

Hence proved.

$\square$

A.3 The pdf of Student's t-distribution

Here, we provide a simple computation of the pdf of the Student's t-distribution.

$$T = \frac{X}{\sqrt{Y/r}}$$

in which $X \sim \mathcal{N}(0, 1)$ and $Y \sim \chi^2(r)$, and they are independent. Thus, we have

$$p(x) \propto e^{-x^2/2}$$
$$p(y) \propto y^{r/2-1} \cdot e^{-y/2}$$

The random variable t follows the equation $t = \frac{x}{\sqrt{y/r}}$.

Lemma A.4. *Since then we want to prove that*

$$p(t) \propto (1 + \frac{t^2}{r})^{-\frac{r+1}{2}} \tag{22}$$

Proof. The joint probability of $p(x, y)$ matches

$$p(x, y) \propto e^{-x^2/2} \cdot y^{r/2-1} \cdot e^{-y/2}$$

And the divergence of $p(x, y)$ is $p(x, y)dxdy$. We can use the variable replacement of

$$y = \frac{x^2}{t^2} \cdot r$$
$$\frac{dy}{dt} \propto \frac{x^2}{t^3}$$

Thus we have the joint probability of $p(x,t)$ matches

$$p(x,t) \propto e^{-x^2/2} \cdot (\frac{x^2}{t^2})^{r/2-1} \cdot e^{-\frac{x^2}{2t^2}r} \cdot \frac{x^2}{t^3}$$

The probability of $p(t)$ can be expressed as

$$p(t) \propto \int_x p(x,t)dx$$

Analysis the expression, we have

$$\begin{aligned}
&p(t) \propto t^{-r-1} \int_x x^r \cdot e^{-\frac{1}{2}(1+\frac{r}{t^2})x^2} dx \\
&p(t) \propto t^{-r-1} \cdot (1+\frac{r}{t^2})^{\frac{-r-1}{2}} \int_z z^r \cdot e^{z^2} dz \\
&p(t) \propto (t^2+r)^{-\frac{r+1}{2}} \\
&p(t) \propto (1+\frac{t^2}{r})^{-\frac{r+1}{2}}
\end{aligned}$$

The process uses the integral of Γ function is constant, and r is constant. □

After that, combining with the following, we should finally have the pdf function.

Lemma A.5. *The values of $t_r(x)$ is positive and the integral is 1.*

$$\int_{-\infty}^{\infty} t_r(x)\,\mathrm{d}x = 1$$

Proof. Consider the variable part of Student's t-distribution

$$f(x) = (1+\frac{x^2}{r})^{-\frac{r+1}{2}}, -\infty < x < \infty$$

use a replacement as following

$$x^2 = \frac{y}{1-y}$$

it is easy to see that $\lim_{y\to 0} x = 0$ and $\lim_{y\to 1} x = \infty$. Additionally, the x^2 is even function. Thus we can write the integral of $f(x)$

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 2\sqrt{r}\int_0^1 (\frac{1}{1-y})^{-\frac{r+1}{2}}\,\mathrm{d}(\frac{y}{1-y})^{\frac{1}{2}}$$

it is not hard to find out that the integral may end up with

$$\sqrt{r}\int_0^1 (1-y)^{\frac{r}{2}-1} y^{\frac{1}{2}-1}\,\mathrm{d}y = \sqrt{r}B(\frac{r}{2},\frac{1}{2})$$

Finally the normalization factor has to be

$$\frac{\Gamma(\frac{r+1}{2})}{\sqrt{r}\Gamma(\frac{r}{2})\Gamma(\frac{1}{2})}$$

which makes the integral of $t_r(x)$ is 1.

□