Statistics

The Normal Distribution and its Applications

Shiu-Sheng Chen

Department of Economics
National Taiwan University

Fall 2019

Section 1

Normal Distributions

Normal Random Variables

- Normal distribution is also called Gaussian distribution, which is named after German mathematician Johann Carl Friedrich Gauss (1777–1855)

  - However, some authors attribute the credit for the discovery of the normal distribution to de Moivre.

Normal Random Variables

Definition (Normal Distribution)

A random variable X has the normal distribution with two parameters μ and σ^2 if X has a continuous distribution with the following pdf:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where $\text{supp}(X) = \{x| -\infty < x < \infty\}$, and $\pi \doteq 3.14159$. It is denoted by $X \sim N(\mu, \sigma^2)$

- Via converting from Cartesian to polar coordinates (google "Gaussian integral")

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = 1$$

Standard Normal Random Variables

Definition (Standard Normal Distribution)

A random variable Z is called a standard normal random variable, if $\mu = 0$ and $\sigma = 1$ with pdf

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

It is denoted by $Z \sim N(0,1)$

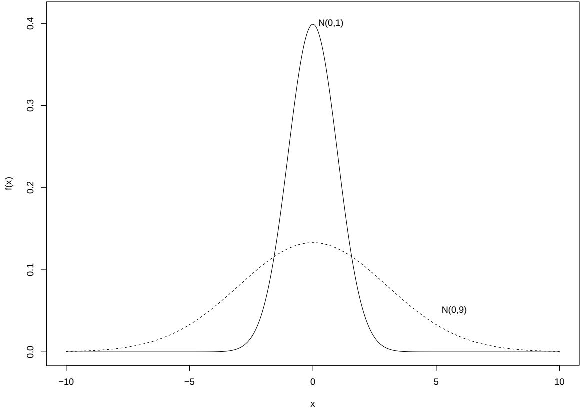
- As a conventional notation, we let ϕ and Φ denote pdf and CDF of a standard normal random variable,

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad \Phi(z) = \int_{-\infty}^{z} \phi(w)dw.$$

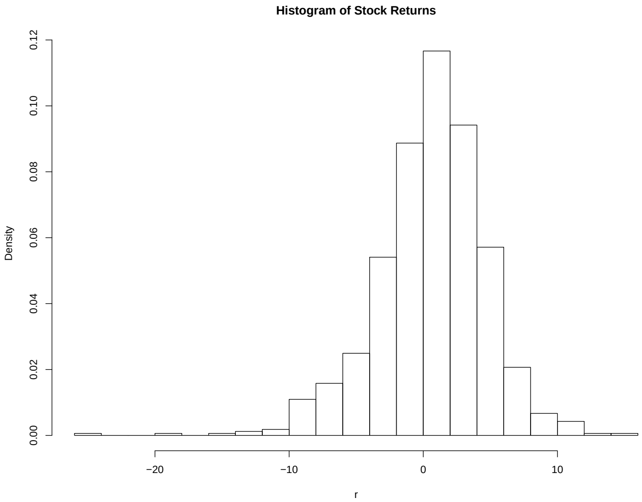
Normal Distributions

Solid line: $N(0,1)$ vs. Dashed line: $N(0,9)$

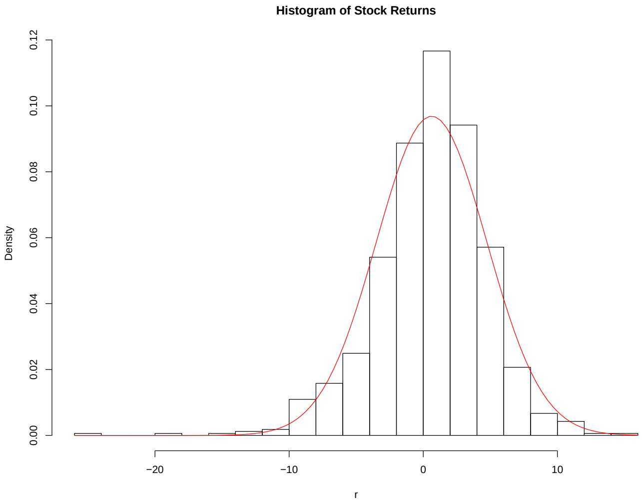
- Skewness $\gamma_3 = 0$; Kurtosis $\gamma_4 = 3$



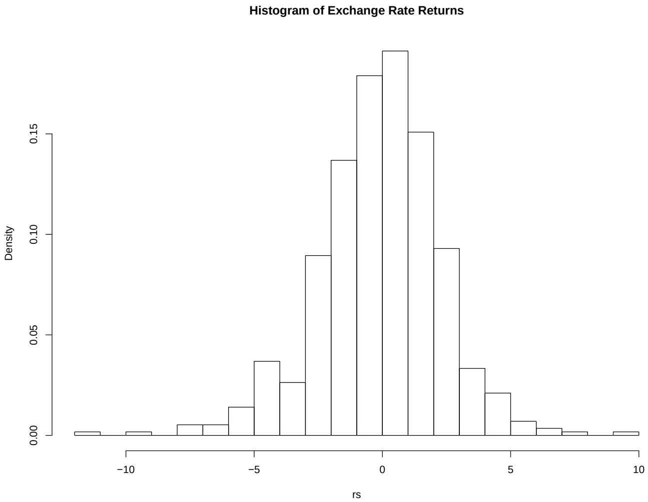
Stock Returns (S&P500)



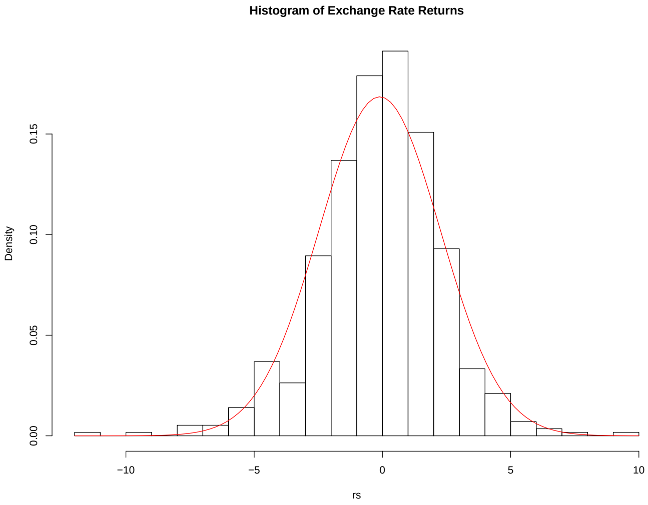
Stock Returns (S&P500)



Exchange Rate Changes (British Pound)



Exchange Rate Changes (British Pound)



Normal vs. Standard Normal

Theorem

Let $Z \sim N(0,1)$ *and* $X = \sigma Z + \mu$, *then*

$$X \sim N(\mu, \sigma^2)$$

- Proof: by CDF method.
- In the same vein, you can show that if $X \sim N(\mu, \sigma^2)$ and $Z = \frac{X-\mu}{\sigma}$, then $Z \sim N(0,1)$.

Moment Generating Function

Theorem (MGF)

Let $Z \sim N(0,1)$, the MGF of Z is

$$M_Z(t) = e^{\frac{1}{2}t^2}$$

- Proof: by definition.
- It follows that if $X \sim N(\mu, \sigma^2)$, the MGF of X is

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

- Hence, $E(X) = \mu$, $Var(X) = \sigma^2$

Properties

Theorem (Invariance Under Linear Transformations)

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

$$aX + b \sim N(a\mu + b, a^2\sigma^2), \quad a \neq 0.$$

- Proof: by MGF
- Example: a portfolio consisting of a stock and a risk-free asset

Properties

Theorem (Sum of I.I.D. Normal Random Variables)

If $\{X_i\}_{i=1}^{n} \sim^{i.i.d.} N(\mu, \sigma^2)$, *and*

$$W = \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_n X_n,$$

then

$$W \sim N\left(\mu \sum_{i=1}^{n} \alpha_i, \sigma^2 \sum_{i=1}^{n} \alpha_i^2\right).$$

- Proof: by MGF
- Consider two special cases:
  - $\alpha_i = 1$ for all i
  - $\alpha_i = \frac{1}{n}$ for all i

Properties

Theorem

If $\{X_i\}_{i=1}^n \sim^{i.i.d.} N(\mu, \sigma^2)$, *then*

$$Y = \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2).$$

If $\{X_i\}_{i=1}^n \sim^{i.i.d.} N(\mu, \sigma^2)$, *then*

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

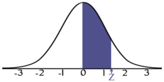
Finding Normal Probabilities

- However, sometimes what we have is the $N(0,1)$ table.
  - When calculating the probability of a normal distribution, we need to transform it to the standard normal distribution.
  - Then calculate $\Phi(a) = P(Z \leq a)$
  - For instance, if $X \sim N(5,16)$,

$$P(X \leq 3) = P\left(\frac{X-5}{4} \leq \frac{3-5}{4}\right) = P(Z \leq -0.5) = \Phi(-0.5)$$

- The following properties will be helpful
  - $P(Z \leq 0) = P(Z \geq 0) = 0.5$
  - $P(Z \leq -a) = P(Z \geq a)$
  - $P(-a \leq Z \leq 0) = P(0 \leq Z \leq a)$
- TA will teach you how to use the $N(0,1)$ table

Z Table



STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and *z* standard deviations above the mean. For example, for *z* = 1.25 the area under the curve between the mean (0) and *z* is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

Using Normal Probabilities to Find Quantiles

- Example: Suppose the final grade, X is normally distributed with mean 70 and standard deviation 10. The instructor wants to give 10% of the class an A+. What cutoff should the instructor use to determine who gets an A+?
- Clearly, $X \sim N(70, 100)$, and we want to find the constant q such that $P(X > q) = 0.10$ or $P(X \leq q) = 0.90$.

Using Z Table

- Find q such that

$$0.90 = P(X \leq q) = P\left(\frac{X-70}{10} \leq \frac{q-70}{10}\right) = P\left(Z \leq \frac{q-70}{10}\right)$$

- According to the Z Table, $P(Z \leq 1.28) = 0.90$, we have

$$\frac{q-70}{10} = 1.28$$

That is,

$$c = 82.8$$

Example: Mean-Variance Utility

- Suppose that the utility function from wealth W is given by

$$U(W) = c - e^{-bW}, \quad b > 0$$

  - This utility function is increasing and concave

$$U'(W) = be^{-bW} > 0, \quad U''(W) = -b^2 e^{-bW} < 0$$

- We further assume that $W \sim N(\mu, \sigma^2)$, then

$$E[U(W)] = c - E\left[e^{-bW}\right] = c - e^{-b(\mu - \frac{b}{2}\sigma^2)} = g(\mu, \sigma^2)$$

- Hence,

$$\frac{\partial E[U(W)]}{\partial \mu} = g_\mu > 0, \qquad \frac{\partial E[U(W)]}{\partial (\sigma^2)} = g_{\sigma^2} < 0$$

Bivariate Normal Random Variables

Definition (Bivariate Normal Distribution)

X and Y are bivariate normal distributed if the joint pdf is

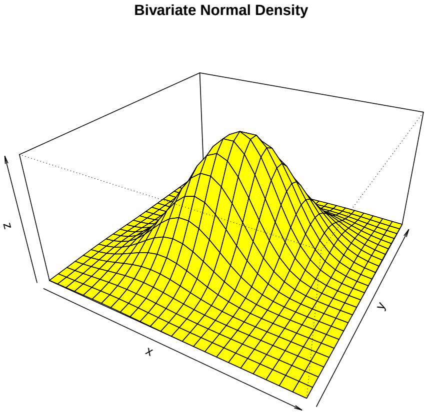
$$f_{XY}(x,y) = \frac{e^{\varpi}}{2\pi\sqrt{(1-\rho^2)}\sigma_X\sigma_Y}$$

where

$$\varpi = -\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]$$

Bivariate Normal Random Variables

- It is denoted by $\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}\right)$



Independent vs. Uncorrelated

Theorem

Given that X and Y have a bivariate normal distribution. X and Y are independent if and only if

$$Cov(X, Y) = 0.$$

- Proof: we only need to show the "if" part. Clearly, when $Cov(X, Y) = 0$, which implies $\rho = 0$, it can be shown that

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

Symmetry vs. Asymmetry

- Normal distribution is a symmetric distribution.
- In some instances, we require a skewed distribution to characterize the data.
- For example,
  - the size of insurance claims
  - time-until-default (survival time)
- We thus introduce a random variable called Chi-square random variable to capture the asymmetrical characteristics.

Section 3

Chi-square Distribution

Chi-square Distribution

Definition (Chi-square Random Variables)

A random variable X has the Chi-square distribution if the pdf is

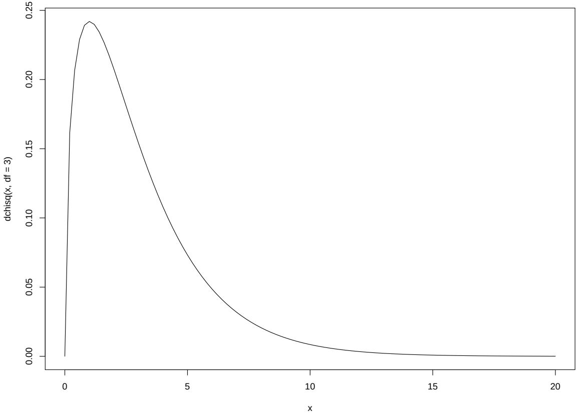
$$f(x) = \frac{x^{\frac{k}{2}-1}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}e^{-\frac{1}{2}x}, \ \ \text{supp}(X) = \{x|0 < x < \infty\}$$

where k is a positive integer called the degree of freedom.

- $\Gamma(\cdot)$ is called a Gamma function:

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

Chi-square distributions ($k = 3$)



Chi-square Distribution

Theorem

The MGF of a Chi-square random variable is

$$M_X(t) = \left(\frac{1}{1-2t}\right)^{\frac{k}{2}}$$

- Proof: by the definition of MGF, and let $y = (1-2t)x$.
- Hence,

$$E(X) = k, \quad Var(X) = 2k$$

Chi-square Distribution

Theorem

If $X_i \sim \chi^2(k_i)$ for $i = 1, 2, \ldots n$, and they are independent, then

$$\sum_{i=1}^{n} X_i \sim \chi^2\left(\sum_{i=1}^{n} k_i\right)$$

- That is, the sum $X_1 + X_2 + \cdots + X_n$ has the χ^2 distribution with $k_1 + k_2 + \cdots + k_n$ degrees of freedom.
- Proof: by MGF.

Chi-square Distribution

The following theorem links the normal distribution and Chi-square distribution.

Theorem

Let $Z \sim N(0,1)$. Then the random variable

$$Y = Z^2 \sim \chi^2(1)$$

- Proof:

$$M_{Z^2}(t) = E(e^{tZ^2}) = \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)z^2} dz = \left(\frac{1}{1-2t}\right)^{\frac{1}{2}}$$

Chi-square Distributions

Corollary

Suppose $\{Z_1, Z_2 \ldots, Z_k\} \sim^{i.i.d.} N(0,1)$. *Let* $X = \sum_{i=1}^{k} Z_i^2$. *Then*

$$X \sim \chi^2(k)$$

- Proof: by the above two theorems.
- Degree of freedom: The number of values in the final calculation of a statistic that are free to vary

Section 4

Student's t Distribution

Student's t Distributions

- Recall that the Kurtosis for Normal random variables is 3
- Monthly S&P 500 Stock Returns (1957:1–2013:9):

$$\text{Kurtosis} = 5.51$$

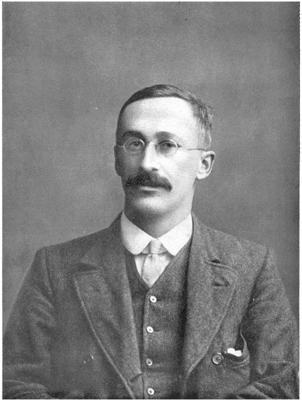
- Daily S&P 500 Stock Returns (1957/1/2–2013/9/30):

$$\text{Kurtosis} = 30.75$$

    - Fat-tailed/Heavy-tailed
- A commonly used heavy-tailed distribution (two-tailed) is the Student's t distribution.

Student's t distributions

- The Student's t distribution was actually published in 1908 by a British statistician, William Sealy Gosset (1876–1937).



William Sealy Gosset

- Gosset, was employed at the Guinness Brewing Co., which forbade its staffs publishing scientific papers due to an earlier paper containing trade secrets.
- To circumvent this restriction, Gosset used the name "Student", and consequently the distribution was named Student's t distribution.

Guinness and the 1908 Biometrika Paper

VOLUME VI | MARCH, 1908 | No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value

Student's t distributions

Definition (Student's t distribution)

If a random variable X has the following pdf

$$\frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})}\frac{1}{\sqrt{k\pi}}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

with support $\text{supp}(X) = \{x| - \infty < x < \infty\}$ and a parameter k, then it is called a Student's t distribution, and denoted by

$$X \sim t(k)$$

- k is called the degrees of freedom.

Student's t distributions

Theorem

Given two independent random variables: $Z \sim N(0,1)$ *and* $W \sim \chi^2(k)$*. Then*

$$U = \frac{Z}{\sqrt{\frac{W}{k}}} \sim t(k)$$

- Proof: beyond the scope of this course.

Student's t distribution

- Given $U \sim t(k)$
- Moments

$$E(U) = 0 \quad \text{when } k > 1,$$

$$Var(U) = E(U^2) = \frac{k}{k-2} \quad \text{when } k > 2.$$

  - Note that given $W \sim \chi^2(k)$,

$$E\left(\frac{1}{W}\right) = \frac{1}{k-2}$$

Student's t distributions

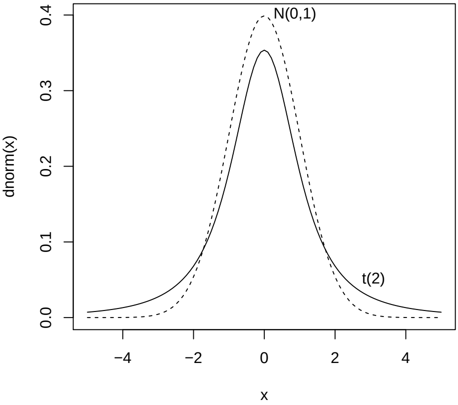
- t distribution is a symmetric distribution.
- Limiting distribution

$$t(k) \longrightarrow N(0,1) \text{ as } k \longrightarrow \infty.$$

- Special case: $k = 1$, $E[t(1)] = \infty - \infty$ (undefined)
    - $t(1)$ is called a standard Cauchy distribution.

Comparison: $t(2)$ vs. N(0,1)

Clearly, the Student's t distribution has a fat tail.



Section 5

F Distributions

F Distribution

Definition (F Distribution)

Random X has the F distribution with n_1 and n_2 degrees of freedom if the probability density function is

$$\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}$$

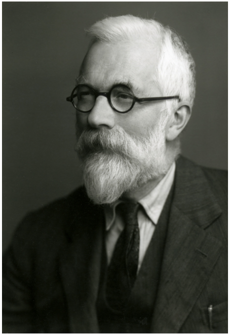
with $\text{supp}(X) = \{x | 0 \le x < \infty\}$. It is denoted by

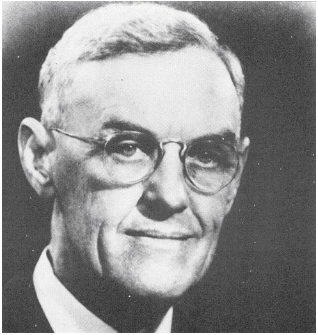
$$X \sim F(n_1, n_2)$$

- It is proposed by R.A. Fisher and George W. Snedecor

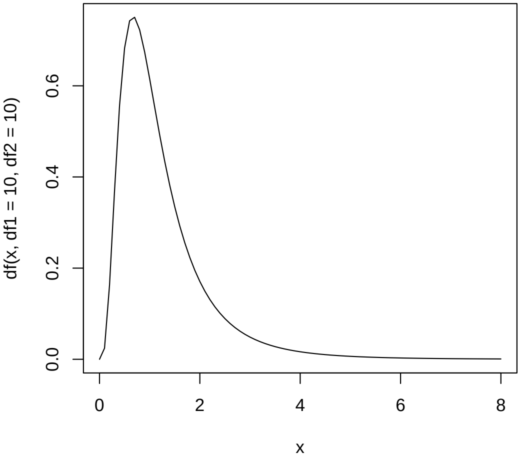
R.A. Fisher and G. W. Snedecor

- Sir Ronald A. Fisher (1890–1962), British statistician, evolutionary biologist, eugenicist, and geneticist
- George W. Snedecor (1881–1974), American mathematician and statistician





F Distribution ($n_1 = 10,\ n_2 = 10$)



F Distribution

Theorem

Let W_1 and W_2 be independent Chi-square random variables:

$$W_1 \sim \chi^2(n_1), \text{ and } W_2 \sim \chi^2(n_2),$$

then

$$X = \frac{W_1/n_1}{W_2/n_2} \sim F(n_1, n_2)$$

- Proof: beyond the scope of this course.

F Distribution

Theorem

If $t \sim t(k)$, then

$$t^2 \sim F(1, k)$$