Econ 136 Review Handout 2: Binomial, Normal and t- Distributions

Edward Vytlacil, Yale University

January 16, 2024

This handout reviews Bernoulli and Binomial distributions, and connects them to indicator/dummy variables. It also reviews the normal distribution, the distribution of the absolute value of a normal random variable, the t-distribution, and the multivariate normal distribution. We will use these distributions extensively for inference.

Contents

Bernoulli Distribution

EXAMPLE 2 OF HANDOUT 1 considered the random variable for a weighted coin toss, with $X = 1$ if heads and $X = 0$ if tails, and $\Pr[X = 1] = p$, $\Pr[X = 0] = 1 - p$. Such a random variable is called a Bernoulli random variable, and its distribution is called the Bernoulli distribution:

Definition 1: Bernoulli Random Variable

A random variable X is distributed *Bernoulli(p)*, $X \sim \text{Bernoulli}(p)$, if:

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

For a Bernoulli random variable, we call $X = 1$ a "*success*" and $X = 0$ a "*failure*".

The terminology of "success" and "failure" need not correspond to good and bad outcomes. For example, $X = 1$ might indicate that a patient dies and $X = 0$ that the patient lives, without implying that death is a good outcome. Likewise, a machine failing or a dam failing might be a "success" despite the tension in the terminology.

While we motivated the Bernoulli distribution using a coin toss, it is the distribution for any *binary variable* taking the values 0 and 1, and is how we encode as a variable yes-no and true-false answers. It is often productive in probability theory (as in economics, as in coding) to abstract from details. Thus, instead of trying to separately consider random variables for whether an individual is employed or not, or a person tests positive for a virus or not, or a machine fails or not, or a dam bursts or not, we instead study Bernoulli random variables and think of hypothetical experiments involving weighted coin flips with our abstract model of weighted coin flips describing any yes-no, true-false type variable, just with different probabilities of "success" or "failure."

In Handout 1, Example 2, we already derived the expectation and variance of a Bernoulli random variable, which we restate here in the following theorem:

Theorem 1: Expected Value and Variance of a Bernoulli r.v.

Suppose $X \sim \text{Bernoulli}(p)$. Then

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

Logical Indicator Function and Indicator Variables

We will often refer to variables representing whether some event is true or not as indicator variables, where the variable equals 1 if the event is true, and equals 0 otherwise. Indicator variables are also called dummy variables. We will often construct indicator variables using logical indicator functions, defined as follows:

Definition 2: Logical Indicator Function

Let $\mathbb{1}$ denote the *logical indicator function*, defined for any statement (event) A as

$$\mathbb{1}[A] = \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{if } A \text{ is false.} \end{cases}$$

We define indicator variables using logical indicator functions:

Definition 3: Indicator Variables

X is an *indicator variable*, also called a *dummy variable*, if, for some given event A, $X = \mathbb{1}[A]$.

If X is an indicator variable for some event A, $X = \mathbb{1}[A]$, then X equals 1 when A is true and equals 0 when A is false. Thus, the indicator variable $X = \mathbb{1}[A]$ is a Bernoulli(p) random variable with $p = \Pr[A]$. For example, we might define $X = \mathbb{1}[\text{employed}]$ as a indicator variable for the event that an individual is employed, which is distributed Bernoulli(p) where p equals probability of being employed. From Theorem 1, we see that the expected value of a indicator variable is the probability that the event occurs, and that the variance of an indicator variable is relatively large when the probability of the event is close to a half and is small when the probability of the event is close to 0 or 1.

We will also often use logical indicator functions to construct indicator variables from some other non-binary random variable, as illustrated by the following examples.

Example 1 (Positive Returns). *Let r_A denote the return on an asset, and define X by $X = \mathbb{1}[r_A > 0]$. Then X is a indicator variable for the asset having a positive return and $\mathbb{E}[X] = \Pr[r_A > 0]$.*

Example 2 (Schooling). *Let S denote an individual's years of schooling. Then $X = \mathbb{1}[S \geq 12]$ is a indicator variable for having completed high school, and $\mathbb{E}[X] = \Pr[S \geq 12]$.*

Dummy variables are pervasive in econometrics, representing, for example, an individual working or not working, having graduated from college or not, being retired or not, being married or not, having children or not, having positive savings or not, having investments in the stock market or not, and so forth.

Binomial Distribution

CONSIDER FLIPPING A (POSSIBLY WEIGHTED) COIN n times, each flip independent of every other, and counting the number of heads. More formally, suppose $X_1, ..., X_n$ denotes i.i.d. Bernoulli(p) random variables so that $S_n = \sum_{i=1}^n X_i$ is number of successes in the n tosses. Then S_n is called a Binomial random variable, and its distribution is called the Binomial(n,p) distribution:

Definition 4: Binomial Distribution

Suppose $X_1, ..., X_n$ are i.i.d. Bernoulli(p) random variables, and let $S_n = \sum_{i=1}^n X_i$. We will refer to each X_i as a Binomial trial, and to S_n as the number of successes in n trials. Then S_n is distributed *Binomial(n, p)*, $S_n \sim \text{Binomial}(n, p)$, with

$$\Pr[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k} \tag{1}$$

for $k = 0, ..., n$, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{2}$$

By their definitions, a Bernoulli(p) random variable is the same as a Binomial(1,p) random variable..

There is a simple intuition for Equation 1. By independence of the Bernoulli trials, any given sequence with k successes (and thus $n-k$ failures) has probability $p^k(1-p)^{n-k}$. For a sequence of size n, there are $\binom{n}{k}$ sequences that result in k succeses. Equation 1 then follows.

The Binomial(n,p) is the number of "successes" of n independent Binomial trials. We will often abstract from any particular context and think of a hypothetical experiment of flipping a weighted coin n times, each flip independent of every other flip, and counting the number of heads, but that hypothetical experiment is a model of any true-false, yes-no outcomes, for example, number of individuals in our sample who are employed, number of investors who invest in the stock market, number of patients who survive, or number of subjects who test positive for an illness as long as the underlying events are independent and have the same probability of success.

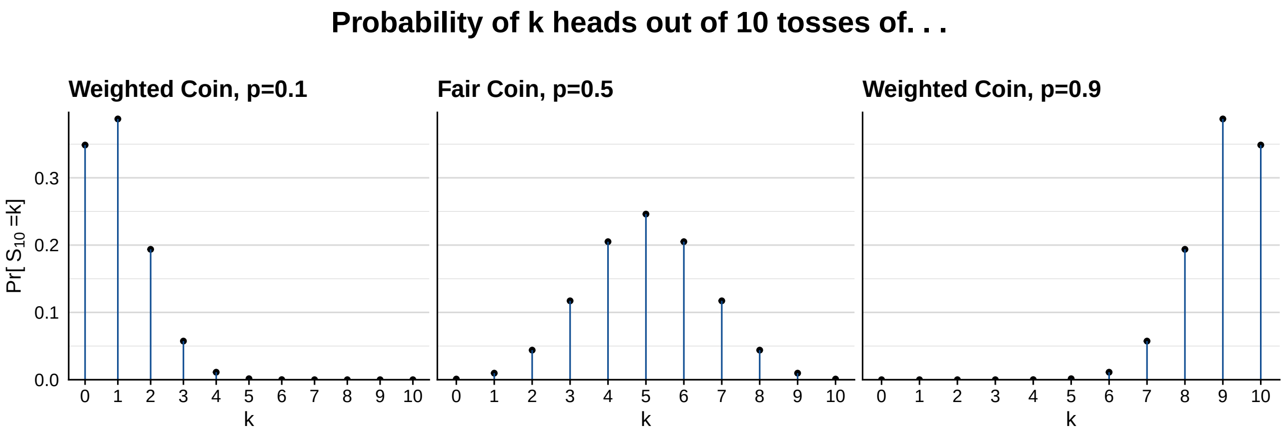


Figure 1: Plotting $\Pr[S_{10} = k] = \binom{10}{k} p^k (1-p)^{10-k}$ for $p = 0.2, 0.5$, and 0.8.

Using rules for expectations and variances (see Handout 1, Theorem 5 and 7, and Remarks 5 and 10) we have the following result:

Theorem 2: Expected Value and Variance of a Bernoulli r.v.

Suppose $S_n \sim \text{Binomial}(n, p)$. Then

$$\mathbb{E}[S_n] = n \cdot p$$
$$\text{Var}[S_n] = n \cdot p \cdot (1-p)$$

The result of Theorem 2 is intuitive. For example, consider flipping a possibly weighted coin ten times. If it is a fair coin, the expected number of heads is 5 heads out of 10. If the probability of heads is 0.2, then the expected number of heads is 2 heads out of 10. Note that our use of the word "expected" is different from the normal dictionary use of the word. For example, if you flip a fair coin 3 times, your expected number of heads is 1.5 heads out of 3, even though you will never flip exactly 1.5 heads out of 3.

Remark 1 (Sample Mean of Independent Bernoulli Trials). *Let* $\bar{X}_n =$

$\frac{1}{n}\sum_{i=1}^{n} X_i$, *the sample mean of the n Bernoulli trials which is equivalent to the fraction of the n Bernoulli trials that are a success. For example, if* X_i *is an indicator variable for being employed, then* $\bar{X}_n$ *is the fraction of the sample that is employed. Since* $\bar{X}_n = S_n/n$, *we can use Theorem 1 along with Theorem 8 of Handout 1 to obtain*

$$\mathbb{E}[\bar{X}_n] = p.$$
$$Var[\bar{X}_n] = p\cdot(1-p)/n.$$

Furthermore, we can obtain the (exact) distribution of the sample mean of n i.i.d. Bernoulli(p) random variables from the Binomial(n,p) distribution:

$$\Pr[\bar{X}_n = t] = \Pr[S_n = n\cdot t] \text{ where } S_n \sim Binomial(n,p). \tag{3}$$

While we can also obtain a normal approximation to (standardized) $\bar{X}_n$ *as n going to infinity using the Central Limit Theorem (CLT), the normal approximation from the CLT will be poor when n is small or p is close to* 0 *or* 1, *while the distribution for* $\bar{X}_n$ *from equation* (3) *is exact for any size n.*

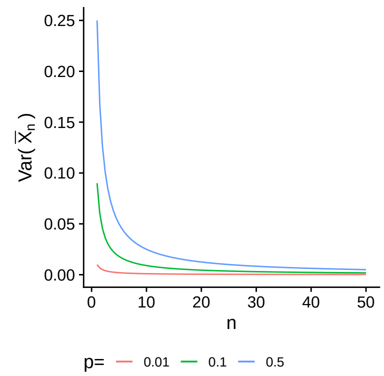


Figure 2: Var($\bar{X}_n$) for n independent Bernoulli(p) trials with p=0.01, 0.10, and 0.50 .

*Binomial Distribution in **R***

We will typically not work with equation (1) directly, but rather use **R** to calculate probabilities for the binomial distribution.

R Functions for Binomial Distribution

Function	Returns
`dbinom(k, n, p)`	$\Pr[S_n = k]$ for $S_n \sim$ Binomial(n,p).
`pbinom(k, n, p)`	$\Pr[S_n \le k]$ for $S_n \sim$ Binomial(n,p).
`qbinom(q, n, p)`	qth quantile of Binomial(n,p)
`rbinom(m, n, p)`	m random draw of S_n, $S_n \sim$ Binomial(n,p)

$\Pr[S_n > k] = 1 - \Pr[S_n \le k]$, and thus we can compute $\Pr[S_n > k]$ by `1-pbinom(k,n,p)`. Alternatively, we can use `pbinom(k,n,p,lower.tail=FALSE)`.

```
> # Example: 10 tosses of fair coin,
>
> # prob of 8 heads out of 10
> dbinom(8,10,0.5)
[1] 0.04394531
> # prob of 8 or fewer heads
> pbinom(8,10,0.5)
[1] 0.9892578
> # prob of more then 8 heads
> # (9 or more)
> 1-pbinom(8,10,0.5)
[1] 0.01074219
```

```
> # Simulating 10 flips of a fair coin
> rbinom(10,1,.5)
 [1] 1 0 1 0 1 1 1 0 0 0
> # another 10 flips of a fair coin
> rbinom(10,1,.5)
 [1] 1 1 0 0 1 0 1 1 0 1
> # Simulating sum of 10 flips
> rbinom(1,10,.5)
[1] 7
> # Simulating sum of 10 flips, 5 times
> rbinom(5,10,.5)
[1] 4 2 7 6 5
```

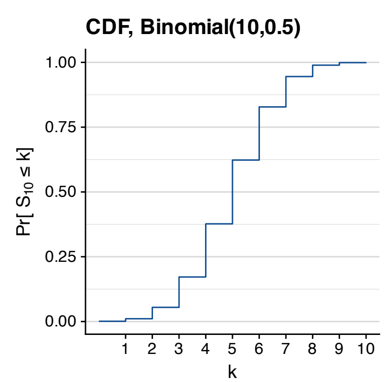


Figure 3: Plotting `pbinom(k, 10, 0.5)`

Normal Distribution

We have thus far considered the Bernoulli(p) and Binomial(n,p) distributions, and we have that many random variables of interest are exactly distributed Bernoulli(p) or Binomial(n,p). We now consider the normal distribution, which is different in that we rarely believe that random variables in practice are exactly normally distributed. However, we often believe that random variables are approximately normally distributed, and normal approximations plays a critical role in large samples due to the Central Limit Theorem.

Definition 5: Normal Distribution

A random variable X is *Normally Distributed* with mean μ and variance σ^2, written $X \sim N(\mu, \sigma^2)$, if the probability density function of X is given by

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

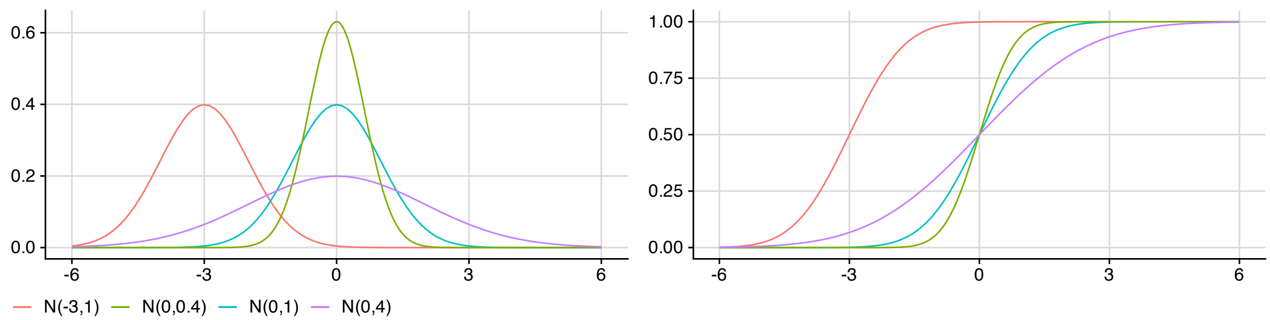


Figure 4: Plotting PDF (left) and CDF (right) of $N(\mu, \sigma^2)$. Note that μ determines where the density is centered and the density is symmetric around μ, while σ^2 determines the spread of the density. Note that — $N(0,1)$ is standard normal.

Particularly important for hypothesis testing will be the normal distribution with $\mu = 0$ and $\sigma^2 = 1$, called the standard normal distribution.

Definition 6: Standard Normal Distribution

A random variable X is distributed *Standard Normal* if $X \sim N(0,1)$. We denote the pdf of a standard normal distribution by $\phi(x)$, so that

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2},$$

and the CDF of a standard normal by $\Phi(x)$, so that

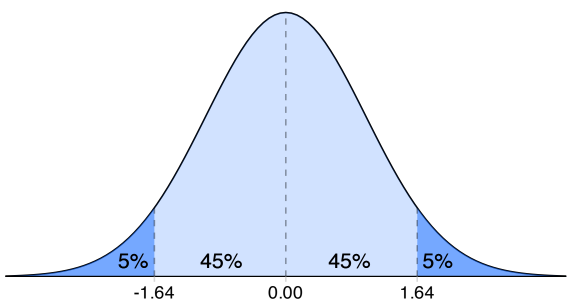
$$\Phi(t) = \Pr[X \leq t] = \int_{-\infty}^{t} \phi(x)dx.$$

Remark 2 (Symmetry of Std. Normal). *A $N(\mu, \sigma^2)$ density is symmetric around μ, so that the standard normal pdf, $\phi(\cdot)$, is symmetric around 0 . The symmetry of $\phi(\cdot)$ around 0 has the following implications, which will be useful for inference:*

1. $\phi(t) = \phi(-t)$ *for all* t;
2. $\Phi(t) = 1 - \Phi(-t)$ *for all* t;
3. $X \sim N(0,1)$ *implies* $-X \sim N(0,1)$.
4. $q_\alpha = -q_{1-\alpha}$ *for all* α, *where* q_α *is the* α *quantile of a standard normal.*

Quantiles of $N(0,1)$

α	q_α
0.01	-2.33
0.025	-1.96
0.05	-1.64
0.95	1.64
0.975	1.96
0.99	2.33



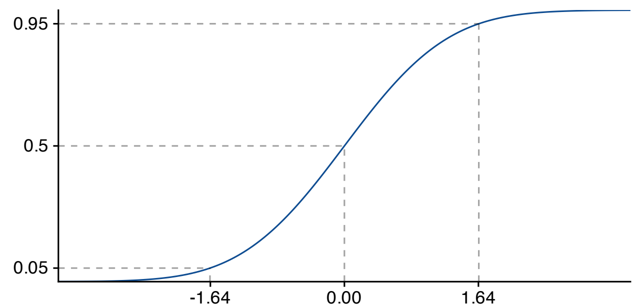


Figure 5: Standard Normal PDF (left) and CDF (right).

We will often work with linear functions of normal random variables.

Theorem 3: Linear Function of Normal

Suppose $X \sim N(\mu, \sigma^2)$. Let a and b denote constants with $b \neq 0$. Then $a + bX \sim N(a + b\mu, b^2\sigma^2)$.

Remark 3 (Converting $N(\mu, \sigma^2)$ r.v. to Std. Normal). *Suppose $X \sim N(\mu, \sigma^2)$. Then, by Theorem 3*

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

Converting a $N(\mu, \sigma^2)$ r.v. to a standard normal r.v. allows us to use Φ to find $\Pr[X \leq t]$ for $X \sim N(\mu, \sigma^2)$. In particular,

$$\begin{aligned} \Pr[X \leq t] =& \Pr\left[\frac{X - \mu}{\sigma} \leq \frac{t - \mu}{\sigma}\right] \\ =& \Phi\left(\frac{t - \mu}{\sigma}\right). \end{aligned}$$

For example, if $X \sim N(1,4)$, then $\Pr[X \leq 2] = \Phi(\frac{2-1}{2}) = \Phi(\frac{1}{2})$.

Consider the sum of independent normal random variables.

Theorem 4: Sum of Independent Normal r.v.'s

Suppose that $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ and X and Y are independent. Let a and b denote constants with $a \neq 0$ or $b \neq 0$. Then

$$aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2).$$

Iterating on Theorem 4 for i.i.d. random variables leads to the following corollary:

Corollary 5: Mean of i.i.d. Normal Random Variables

Suppose $X_1, X_2, ..., X_N$ are i.i.d. with $X_i \sim N(\mu, \sigma^2)$. Then

$$\bar{X}_N \sim N(\mu, \sigma^2/N).$$

Following Remark 3 and applying Corollary 5, we have that, if $X_1, X_2, ..., X_N$ are i.i.d. with $X_i \sim N(\mu, \sigma^2)$, then

$$\frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}} \sim N(0,1),$$

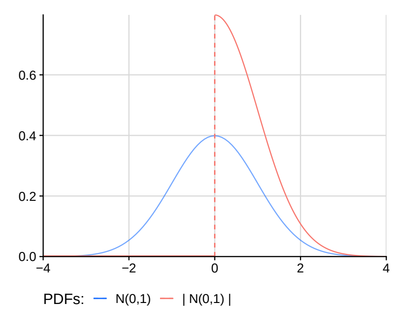
so that

$$\Pr[\bar{X}_N \leq t] = \Phi\left(\frac{t-\mu}{\sigma/\sqrt{N}}\right).$$

Remark 4 (Absolute Value of Std. Normal)**.** *The distribution of the absolute value of a standard normal r.v. is important for inference. Suppose $X \sim N(0,1)$ so that $|X| \sim |N(0,1)|$. Then, using the symmetry of $\phi(\cdot)$,*

- *for $x \geq 0$, the density of $|X|$ at x is $2 \cdot \phi(x)$;*
- *for $x \geq 0$, $\Pr[\,|X| > x\,] = 2 \cdot \Pr[X > x] = 2\,(1 - \Phi(x))$;*
- *the $1-\alpha$ quantile of $|X|$ equals the $1-\alpha/2$ quantile of X.*

For example, if $X \sim N(0,1)$, then $\Pr[X > 1.64] = 1 - \Phi(1.64) = 0.05$, $\Pr[|X| > 1.64] = 2 \cdot (1 - \Phi(1.64)) = 0.10$, and thus 1.64 is the 0.95 quantile of $X \sim N(0,1)$ which is the 0.90 quantile of $|X|$.



$|N(0,\sigma^2)|$ is called the *half-normal* distribution, so that $|N(0,1)|$ is an example of a half-normal distribution .

Normal Distribution in ***R***

We will typically use **R** to calculate probabilities for the normal distribution.

R Functions for Normal Distribution

Function	Returns
`dnorm(x, m, s)`	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-m}{s}\right)^2}$, the $N(m, s^2)$ density evaluated at x,
`pnorm(x, m, s)`	$\Pr[X \leq x]$ for $X \sim N(m, s^2)$.
`qnorm(q, m, s)`	qth quantile of $N(m, s^2)$,
`rnorm(n, m, p)`	n random draw of X, $X \sim N(m, s^2)$.

These functions set $m = 0$ and $s = 1$ by default if their values are not specified. Thus, `dnorm(0.5)` returns the same value as `dnorm(0.5, 0, 1)`.

```
> # 0.05 Quantile of N(0,1)
> qnorm(0.05,0,1)
[1] -1.644854
> # Std. Norm Density at -1.645
> dnorm(-1.6449,0,1)
[1] 0.1031278
> # Prob Std. Norm. Less than -1.645
> pnorm(-1.645,0,1)
[1] 0.04998491
> # Prob Std. Norm. Greater than -1.645
> 1-pnorm(-1.645,0,1)
[1] 0.9500151
> # Default mu=0, sigma=1
> pnorm(-1.645)
[1] 0.04998491
```

```
> # .95 Quantile of N(0,1)
> qnorm(0.95,0,1)
[1] 1.644854
> # Std. Norm Density at 1.645
> dnorm(1.6449,0,1)
[1] 0.1031278
> # Prob Std. Norm. Less than 1.645
> pnorm(1.645,0,1)
[1] 0.9500151
> # Prob Std. Norm. Greater than 1.645
> 1-pnorm(1.645,0,1)
[1] 0.04998491
> # Prob -1.64<Std. Norm.< 1.64
> pnorm(1.645)-pnorm(-1.645)
[1] 0.9000302
```

```
> # Simulating 1 draw from N(0,1)
> rnorm(1,0,1)
[1] 1.207962
> # Another draw, using defaults
> rnorm(1)
[1] -0.4028848
 > # Simulating 3 draws from N(0,1)
> rnorm(3,0,1)
[1]  0.55391765 -0.06191171 -0.30596266
 > # Another 3 draws  from N(0,1)
> rnorm(3)
[1] -0.46665535  0.77996512 -0.08336907
```

```
> # Consider X ~ N(1,4)
> # Pr[X<=2]
> pnorm(2,1,2)
[1] 0.6914625
> # Using Remark 3
> # Pr[X<=2]
> pnorm(0.5,0,1)
[1] 0.6914625
> # Simulating 3 draws from N(1,4)
> # Using Theorem 3
> 1 + 2 * rnorm(3)
[1]  0.7841654 -1.3307927  5.5379119
```

t-Distribution

A family of distributions closely related to the standard normal distribution is the t-distribution, which will play a key role in inference.

Definition 7: t-distribution

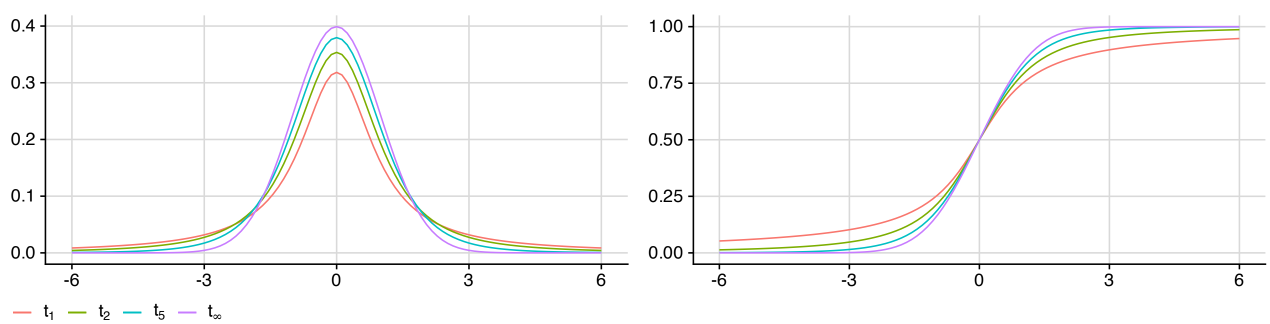
A random variable X has a *t-distribution* with ν degrees of freedom, written $X \sim t_\nu$, if its probability density function is given by

$$f_X(x) = \frac{\Gamma(\nu+1)/2}{\Gamma(\nu/2)\sqrt{\nu\pi}}(1+\frac{x^2}{\nu})^{-(\nu+1)/2}, \tag{4}$$

where ν is a positive integer and Γ is the gamma function.

In this course, you do not need to remember the formula for the normal or t-densities, and we will never work with them directly but rather use **R** when we need to evaluate them.

The t-distribution is a family of distributions indexed by the parameter ν, called the degrees of freedom. Like the standard normal distribution, the t-distribution is symmetric around zero and bell-shaped. However, the tails of the t-distribution are heavier than those of the normal distribution, with how much heavier depending on the parameter ν. When $\nu = 1$, the distribution is called the *cauchy distribution* and has much heavier tails and very different properties than a standard normal distribution. The larger is ν, the thinner the tails, and the closer the t_ν distribution is to a standard normal distribution. As ν goes to infinity, the t_ν distribution approaches t_∞, which is the $N(0,1)$ distribution.

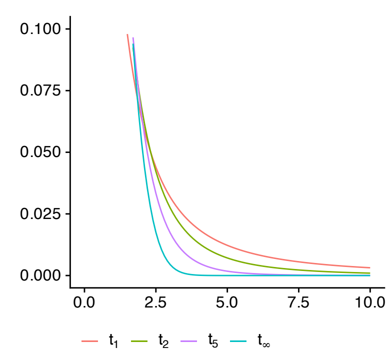


The heavier tails of a t-distribution results in extreme values being more likely for a t-distribution than for a standard normal distribution, especially for ν small. In the extreme case of $\nu = 1$, i.e., a

Examples of t_ν distribution

ν	$\mathbb{E}[X]$	$\text{Var}(X)$	$\Pr[\lvert X\rvert > 3]$	$\Pr[\lvert X\rvert > 5]$	name
1	Does not exist	Does not exist	0.205	0.126	Cauchy
2	0	Does not exist	0.095	0.038	
3	0	3	0.058	0.015	
4	0	2	0.040	0.007	
5	0	$1\frac{2}{3}$	0.030	0.004	
∞	0	1	0.003	0.000	Std. Normal

Figure 6: Plotting the t_ν PDF (above left), CDF (above right), and right tail of PDF (below). Note that — t_1 is the Cauchy density, and — t_∞ is the $N(0,1)$ density.



cauchy distribution, the tails are so heavy that that $\mathbb{E}[X]$ and $\text{Var}(X)$

do not exist. For a t_2 distribution, $\mathbb{E}[X]$ does exists and equals 0, but $\text{Var}(X)$ does not exist. For a t_ν distribution with $\nu \geq 3$, then $\mathbb{E}[X] = 0$ and $\text{Var}(X) = \frac{\nu}{\nu-2}$. For a t_∞ distribution, i.e., a $N(0,1)$ distribution, $\mathbb{E}[X] = 0$ and $\text{Var}(X) = 1$.

A normal distribution can take any value on the whole real line, and thus can take extreme values. However, the tails of a normal distribution go to zero so quickly (i.e., are so thin) that one can essentially ignore the possibility of extreme values with a normal distribution. For example, if $X \sim N(0,1)$, then $\Pr[|X| > 5] = 0.0000006$. In contrast, for a t-distribution with small ν, one cannot ignore the extreme values. For example, if $X \sim t_1$ (Cauchy), $\Pr[|X| > 5] = 0.126$ and X takes extreme values so often that $\mathbb{E}[X]$ and $\text{Var}(X)$ don't exist. Researchers sometimes use the t-distribution, especially Cauchy, when modeling variables where it is important to account for the variables taking extreme values. It is often used in physics, but is also sometimes used to model asset returns and study financial risk in the context of returns that take extreme values with too high of a probability to be ignored.

The Cauchy distribution is also used in probability theory to provide an example where standard results do not hold because $\mathbb{E}(X)$ and $\text{Var}(X)$ do not exist. For example, the LLN and CLT do not hold for the sample mean of i.i.d. draws from a Cauchy. In fact, the sample mean of n i.i.d. Cauchy random variables has the same distribution as one Cauchy random variable.

However, the most common use of a t-distribution in statistics is to model the distribution of the *studentized* mean (also called the *t-statistic* or *t-ratio*) when sampling from a normal distribution. Suppose $X_1, X_2, ..., X_n$ are i.i.d. with $X_i \sim N(\mu, \sigma^2)$. Then it follows from Corollary 5 that $\bar{X}_n \sim N(\mu, \sigma^2/n)$ so that $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, where it is important to note that the expressions are dividing by the true, population σ which is generally unknown. Suppose we had a value of μ that we hypothesized as the true value, but don't know the value of σ^2, so that we need to estimate it. Define the studentized mean as

$$T_n = \frac{\bar{X}_n - \mu}{s_n/\sqrt{n}} \tag{5}$$

where

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2. \tag{6}$$

Then $T_n \sim t_{n-1}$, as stated in the following theorem.

Theorem 6: Studentized Mean of i.i.d. Normal r.v.s

Suppose $X_1, X_2, ..., X_N$ are i.i.d. with $X_i \sim N(\mu, \sigma^2)$. Let T_n and s_n^2 be defined by equations (5) and (6). Then $T_n \sim t_{n-1}$.

An implication of Theorem 6 is that when n is small, the Studentized statistic $\frac{\bar{X}_n - \mu}{s_n/\sqrt{n}}$ will be far from normal and will take extreme values relatively frequently because we are dividing by s_n instead of σ. On the other hand, when n is large, the distribution of $\frac{\bar{X}_n - \mu}{s_n/\sqrt{n}}$ will be close to normal and that we are substituting s_n for σ will be unimportant.

Student-t Distribution in ***R***

We will typically use **R** to calculate probabilities for the Student-t distribution.

R Functions for Student-t Distribution

Function	Returns
`dt(x, v)`	the PDF of a t_v density evaluated at x,
`pt(x, v)`	$\Pr[X \leq x]$ for $X \sim t_v$.
`qnorm(q, v)`	qth quantile of t_v,
`rt(n, v)`	n random draw of X, $X \sim t_v$.

```
> # 0.05 Quantile of Cauchy
> qt(0.05,1)
[1] -6.313752
> # 0.05 Quantile of t_5
> qt(0.05,5)
[1] -2.015048
> # 0.05 Quantile of t_infinity
> qnorm(0.05)
[1] -1.644854
> # density of Cauchy at -5
> dt(-5,1)
[1] 0.01224269
> # density of t_5 at -5
> dt(-5,5)
[1] 0.001757438
> # density ot t_infinity at -5
> dnorm(-5)
[1] 0.00000148672
```

```
> # Prob t_1 between -5 and 5
> pt(5,1)-pt(-5,1)
[1] 0.8743341
> # Prob  | t_1 | >5
> 1 - (pt(5,1)-pt(-5,1))
[1] 0.1256659
> # Prob t_5 between -5 and 5
> pt(5,5)-pt(-5,5)
[1] 0.9958953>
# Prob  | t_5 | >5
> 1 - (pt(5,5)-pt(-5,5))
[1] 0.004104716
> # Prob t_infinity between -5 and 5
> pnorm(5)-pnorm(-5)
[1] 0.9999994
> # Prob  | t_infinity | >5  3
> 1 - (pnorm(5)-pnorm(-5))
[1] 0.0000005733031
```

```
> # Simulating 6 draw from Cauchy
 > rt(6,1)
[1] -0.1055119 -3.5727334 -0.1666175
     14.4183067  2.8204201 -0.1955553
> # Another 6 draws from Cauchy
> rt(6,1)
  [1] -0.4865291 -0.4264724  1.2611932
       1.0142861 -0.6185245 -0.4247870
 > # Simulating 6 draws from t_5
> rt(3,5)
[1] -0.7905094 -0.2588564 -0.3107277
     1.3989390  0.7775500 -0.4130591
```

```
> # another 6 draws from t_5
> rt(3,5)
[1]  -2.0170308  1.0184921  1.7427500
      0.3218333  1.1276689 -0.1284361
> # Simulating 6 draws  from t_infinity
> rnorm(6)
[1] -0.6989059  0.4868481 -0.8626596
     1.5333999  0.4064118  0.1865867
> # Another 6 draws  from t_infinity
> rnorm(6)
[1] -1.5115780  2.0045310 -1.7778798
    -0.8635079 -0.1826787  0.2622432
```

Note that many draws from a Cauchy look very similar to draws from t_5 or from $N(0,1)$, though once in a while the Cauchy takes extreme values.

Multivariate Normal Distribution

We have thus far considered distributions for scalar random variables. We now consider the multivariate normal distribution, which generalizes the definition of normal distribution. The multivariate normal distribution will play a critical role in estimation theory and inference.

Definition 8: Multivariate Normal Distribution

A $K-$dimensional random vector $\mathbf{X} = [X_1, ..., X_K]^T$ is *Multivariate Normally Distributed*, also called *joint normal*, with mean $\boldsymbol{\mu} = [\mu_1, ..., \mu_K]^T$ and invertible variance matrix Σ, written $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, if the probability density function of $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(x_1, ..., x_K) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}}{\sqrt{(2\pi)^K |\Sigma|}}.$$

Remark 5. *The multivariate normal distribution with $K = 2$ is called the bivariate normal distributoin.*

Remark 6. *In general, X_1 and X_2 being uncorrelated does not imply that they are independent. However, if $[X_1, X_2]$ is bivariate normally distributed, then $Cov(X_1, X_2) = 0$ implies that X_1 and X_2 are independent.*

We will often work with linear functions of normal random vectors. A key result is that if $\mathbf{X}$ is joint normal, than any linear combination of $\mathbf{X}$ is normally distributed. The following theorem states this result generalizing Theorem 3:

Theorem 7: Linear Function of Multivariate Normal

Suppose $\boldsymbol{X}$ is a $K-$dimensional random vector with $\mathbf{X} \sim N(\mu, \Sigma)$. Let $\mathbf{a}$ denote an $L \times 1$ constant vector, and $\mathbf{B}$ an $L \times K$ constant matrix, with $L \leq K$. Then the $L-$dimensional random vector $\mathbf{a} + \mathbf{BX}$ is multivariate normally distributed,

$$\boldsymbol{a} + \boldsymbol{BX} \sim N(\boldsymbol{a} + \boldsymbol{B\mu}, \boldsymbol{B}\Sigma\boldsymbol{B}^T).$$

Remark 7 (Subvectors of Multivariate Normal Random Vectors)**.** *By Theorem 7, $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ implies each subvector of $\mathbf{X}$ is normally distributed. For example, if $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$ with*

$$\mathbf{X} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{pmatrix}\right),$$

then $\mathbf{X}_1 \sim N(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathbf{X}_2 \sim N(\boldsymbol{\mu}_2, \Sigma_2)$. Thus, for example, if $\mathbf{X}$ has a multivariate normal distribution, each random variable in $\mathbf{X}$ has a univariate normal distribution.

Remark 8. *Following Remark 7, if $[X_1, X_2]$ is bivariate normally distributed, then X_1 and X_2 are both individually normally distributed. In general, the reverse direction does not hold: X_1 being normally distributed and X_2 being normally distributed does not imply that $\mathbf{X} = [X_1, X_2]^T$ is joint normal. However, if X_1 and X_2 are independent, then they being individually normally distributed does imply that $[X_1, X_2]^T$ is joint normal. More generally, if $X_1, X_2, ..., X_K$ are mutually independent random variables, then with each random variable being normally distributed implies that $\mathbf{X} = [X_1, ..., X_K]^T$ is multivariate normally distributed. An implication of this result is that that Theorem 4 and Corollary 5 are special cases of Theorem 7.*

While Definition 8 generalizes the definition of the normal distribution for scalar random variables from Definition 5, the following definition generalizes the definition of standard normal distribution from Definition 6

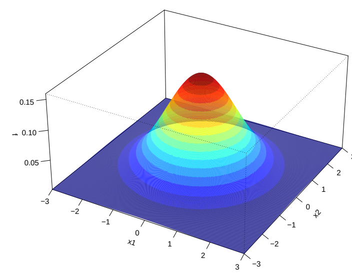


Figure 7: Bivariate Standard Normal Density.

Definition 9: Multivariate Standard Normal Distribution

A random vector $\boldsymbol{X}$ is distributed *Multivariate Standard Normal* if $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_K)$, where $\mathbf{0}$ is a $K \times 1$ vector of zeros and $\mathbf{I}_K$ is the $K \times K$ identity matrix, in which case the corresponding density is

$$f_{\boldsymbol{X}}(x_1, ..., x_K) = \frac{e^{-\frac{1}{2}\mathbf{x}^T\mathbf{x}}}{\sqrt{(2\pi)^K}}.$$

Remark 9. *. Suppose $\boldsymbol{X}$ is distributed multivariate standard normal, $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_K)$. Then, by by Theorem 7, $\boldsymbol{\mu} + B\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ where $\Sigma = BB^T$.*

Summary

Important Definitions

Def 1: $X \sim$ **Bernoulli(p)** if $\Pr[X = 1] = p$, $\Pr[X = 0] = 1 - p$. We call $X = 1$ a "success" and $X = 0$ a "failure".

Def 2: $\mathbb{1}$ denotes the **logical indicator** function where, for any event A, $\mathbb{1}[A] = 1$ if A is true and $\mathbb{1}[A] = 0$ if A is false.

Def 3: X is an **indicator variable**, also called a **dummy variable**, if, for some given event A, $X = \mathbb{1}[A]$.

Def 4: If $X_1, ..., X_n$ are i.i.d. Bernoulli(p) r.v.'s, then $S_n = \sum_{i=1}^n X_i$ is distributed **Binomial(n,p)** with $\Pr[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k}$.

Def 5: X is **Normally Distributed** with mean μ and variance σ^2, written $\boldsymbol{X \sim N(\mu, \sigma^2)}$, if the pdf of X is given by

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

Def 6: X is distributed **Standard Normal** if $X \sim N(0, 1)$. We denote the standard normal pdf by. $\boldsymbol{\phi}(\cdot)$ and the standard normal CDF by $\boldsymbol{\Phi}(\cdot)$.

Def 7: X is distributed according to a **t-distribution** with ν degrees of freedom, written $X \sim t_\nu$, if it's pdf is given by (4).

- When $\nu = 1$, the distribution is called the **Cauchy** distribution.
- As $\nu \to \infty$, the distribution approaches t_∞, the standard normal distribution.

Def 8: $\mathbf{X}$ is distributed *Multivariate Normal*, also called *joint normal*, written $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, if the pdf of $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(\boldsymbol{x}) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}}{\sqrt{(2\pi)^K |\Sigma|}}.$$

Def 9: $\mathbf{X}$ is distributed *Multivariate Standard Normal* if $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_K)$.

Important Results

Thm 1: If $X \sim$ Bernoulli(p) then $\mathbb{E}[X] = p$,
$\text{Var}(X) = p \cdot (1 - p)$.

Thm 2: If $S_n \sim$ Binomial(n,p), then $\mathbb{E}[S_n] = n \cdot p$,
$\text{Var}[S_n] = n \cdot p \cdot (1 - p)$.

Thm 3: If $X \sim N(\mu, \sigma^2)$, and a and b denote constants with $b \neq 0$, then $a + bX \sim N(a + b\mu, b^2\sigma^2)$.

Thm 4: If $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$, X and Y are independent, and a and b denote constants with $a \neq 0$ or $b \neq 0$, then
$aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$.

Cor 5: Suppose $X_1, X_2, ..., X_N$ are i.i.d. with $X_i \sim N(\mu, \sigma^2)$. Then $\bar{X}_N \sim N(\mu, \sigma^2/N)$ so that $\frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}} \sim N(0, 1)$.

Thm 6: Suppose $X_1, X_2, ..., X_N$ are i.i.d. with $X_i \sim N(\mu, \sigma^2)$. Then $\frac{\bar{X}_N - \mu}{s_n/\sqrt{N}} \sim t_{n-1}$, where $s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$.

Thm 7: If $\mathbf{X} \sim N(\mu, \Sigma)$, then

$$\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X} \sim N(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\Sigma\boldsymbol{B}^T).$$