Lecture 4: Probability

Jacob M. Montgomery

Quantitative Political Methodology

Lecture 4

Class business

- PROBLEM SET 1 IS DUE RIGHT NOW
- Problem set 2 will be distributed today via the syllabus

Facebook and survey

- Sign up for our Facebook group: https://www.facebook.com/groups/1071702902960687/
- Take the class survey! Can't assign teams until you all do. https://wustl.az1.qualtrics.com/jfe/form/SV_6rpSYD3xxmbRe5v

Roadmap

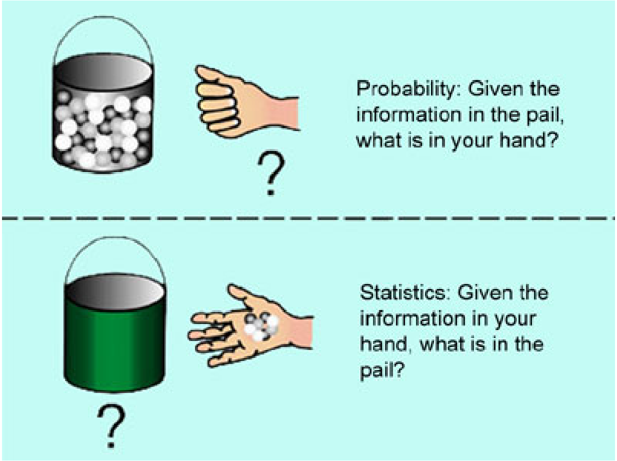
Last time:

- Visualizing data
- Measures of central tendency and spread

This time:

- Understand core concepts of probability
- Understand concept of a “parameter”
- Introduce some probability distributions

Why are we studying this?



Probability defined

Imagine tossing a coin…

- Can you predict the outcome of a single coin toss?

Probability defined

Imagine tossing a coin...

- Can you predict the outcome of a single coin toss?
- Can you predict the *overall* outcome of 100 coin tosses?

Probability defined

Imagine tossing a coin. . .

- Can you predict the outcome of a single coin toss?
- Can you predict the *overall* outcome of 100 coin tosses?

AF p. 73: "For a particular possible outcome for a random phenomenon, the probability of that outcome is the proportion of times that the outcome would occur in a very long sequence of observations."

Example

Imagine you were rolling two six-sided dice.

1. Write down all possible scores.
2. Calculate the probability of each score
   - What is the probability of rolling a 2?

36 possible outcomes for the two dice:

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

How many outcomes will generate a total score of 2?

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

P(roll=2) $= \frac{1}{36} = 0.028$.

Putting this all together

y_k	$Pr(Y = y_k)$
2	$1/36$
3	$2/36$

Putting this all together

y_k	$Pr(Y = y_k)$
2	$1/36$
3	$2/36$
4	$3/36$
5	$4/36$
6	$5/36$
7	$6/36$
8	$5/36$
9	$4/36$
10	$3/36$
11	$2/36$
12	$1/36$

More formal definition

Probability *is the relative frequency of occurrence for some particular outcome if a process is repeated a large number of times under similar conditions*

More formal definition

Probability *is the relative frequency of occurrence for some particular outcome if a process is repeated a large number of times under similar conditions*

- If I flip a coin three times, what is the probability that I will get exactly two heads?
- If I roll two dice, what is the probability of getting a two?
- If I take a random sample of 100 Wash U students, what is the probability that less than 40% of the sample will be male?

Frequency distributions

***Frequency** distribution (discrete probability distribution) is a probability distribution of a discrete variable Y assigns a probability to each possible outcome.*

Frequency distributions

Frequency *distribution (discrete probability distribution) is a probability distribution of a discrete variable Y assigns a probability to each possible outcome.*

- Let $S = \{y_1, y_2, \ldots, y_k\}$ be the set of all possible outcomes,

Frequency distributions

Frequency *distribution (discrete probability distribution) is a probability distribution of a discrete variable Y assigns a probability to each possible outcome.*

- Let $S = \{y_1, y_2, \ldots, y_k\}$ be the set of all possible outcomes, and Y be the realization of the variable.

Frequency distributions

***Frequency** distribution (discrete probability distribution) is a probability distribution of a discrete variable Y assigns a probability to each possible outcome.*

- Let $S = \{y_1, y_2, \ldots, y_k\}$ be the set of all possible outcomes, and Y be the realization of the variable.

- Then, $p(y_k) = Pr(Y = y_k)$, where
- $0 \leq p(y_k) \leq 1 \; \forall \; k$

Frequency distributions

Frequency *distribution (discrete probability distribution) is a probability distribution of a discrete variable Y assigns a probability to each possible outcome.*

- Let $S = \{y_1, y_2, \ldots, y_k\}$ be the set of all possible outcomes, and Y be the realization of the variable.

- Then, $p(y_k) = Pr(Y = y_k)$, where
- $0 \leq p(y_k) \leq 1 \;\forall\; k$
- $$\sum_{k=1}^{K} p(y_k) = 1$$

We already made one of these

y_k	$Pr(Y = y_k)$
2	$1/36$
3	$2/36$
4	$3/36$
5	$4/36$
6	$5/36$
7	$6/36$
8	$5/36$
9	$4/36$
10	$3/36$
11	$2/36$
12	$1/36$

We already made one of these

y_k	$Pr(Y = y_k)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

- $p(y_k) = Pr(Y = y_k)$
- $0 \leq p(y_k) \leq 1 \ \forall \ k$
- $\sum_{k=1}^{K} p(y_k) = 1$

Parameters of distributions

- In probability theory we often wish to identify two important characteristics of distribution.

Parameters of distributions

- In probability theory we often wish to identify two important characteristics of distribution.
  - We wish to know the mean (μ), which is sometimes called the expected value.
  - We wish to know the variance (σ^2)
- NOTE: This is **not** the same as $\bar{x}$ and s^2.

Parameters of distributions

- In probability theory we often wish to identify two important characteristics of distribution.
  - We wish to know the mean (μ), which is sometimes called the expected value.
  - We wish to know the variance (σ^2)
- NOTE: This is **not** the same as $\bar{x}$ and s^2. Why are these Greek letters?

Expected value

y_k	$Pr(Y = y_k)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Mean of probability distribution (**expected value**)

$$\mu = \sum_{k=1}^{K} y_k Pr(Y = y_k)$$

Expected value

y_k	$Pr(Y = y_k)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Mean of probability distribution (**expected value**)

$$\mu = \sum_{k=1}^{K} y_k Pr(Y = y_k) = 2(1/36) + 3(2/36) + \ldots + 12(1/36)$$

Expected value

y_k	$Pr(Y = y_k)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Mean of probability distribution (**expected value**)

$$\mu = \sum_{k=1}^{K} y_k Pr(Y = y_k) = 2(1/36) + 3(2/36) + \ldots + 12(1/36) = 7$$

Expected value

y_k	$Pr(Y = y_k)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Mean of probability distribution (**expected value**)

$$\mu = \sum_{k=1}^{K} y_k Pr(Y = y_k) = 2(1/36) + 3(2/36) + \ldots + 12(1/36) = 7$$

The variance of a distribution

y_k	$Pr(Y = y_k)$
2	$1/36$
3	$2/36$
4	$3/36$
5	$4/36$
6	$5/36$
7	$6/36$
8	$5/36$
9	$4/36$
10	$3/36$
11	$2/36$
12	$1/36$

$$\sigma^2 = E(Y - \mu)^2$$

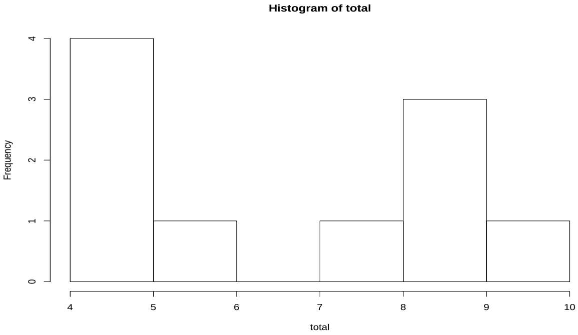
The variance of a distribution

y_k	$Pr(Y = y_k)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

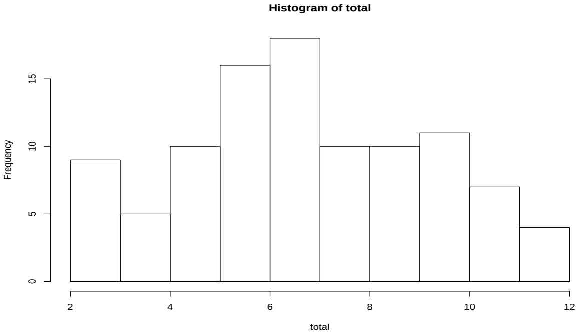
$\sigma^2 = E(Y - \mu)^2$ requires extra calculations

A little simulation

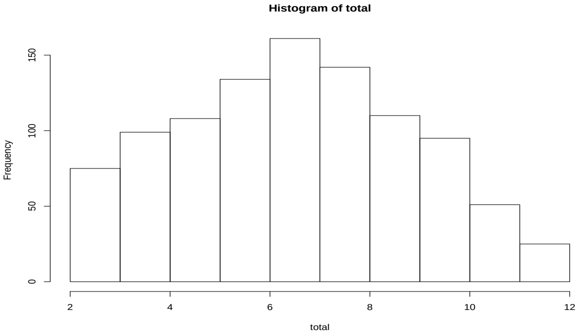
```
posVal<-c(1,2,3,4,5,6)
numRoll<-10
die1<-sample(x = posVal, size=numRoll, replace=TRUE)
die2<-sample(x = posVal, size=numRoll, replace=TRUE)
total<-die1+die2
hist(total)
```



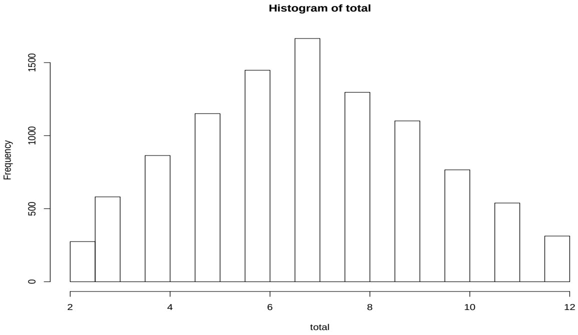
```r
posVal<-c(1,2,3,4,5,6)
numRoll<-100
die1<-sample(x = posVal, size=numRoll, replace=TRUE)
die2<-sample(x = posVal, size=numRoll, replace=TRUE)
total<-die1+die2
hist(total)
```



```r
posVal<-c(1,2,3,4,5,6)
numRoll<-1000
die1<-sample(x = posVal, size=numRoll, replace=TRUE)
die2<-sample(x = posVal, size=numRoll, replace=TRUE)
total<-die1+die2
hist(total)
```



```r
posVal<-c(1,2,3,4,5,6)
numRoll<-10000
die1<-sample(x = posVal, size=numRoll, replace=TRUE)
die2<-sample(x = posVal, size=numRoll, replace=TRUE)
total<-die1+die2
hist(total)
```



End of part 1

- What is a probability?
- What is a frequency distribution?
- What are the two most important parameters for characterizing a distribution?

Example: The Binomial Distribution

Imagine tossing a fair coin in the air three times, where we are interested in the number of heads.

Coin 1	Coin 2	Coin 3	# Heads
H	H	H	3
T	H	H	2
H	T	H	2
T	T	H	1
H	H	T	2
T	H	T	1
H	T	T	1
T	T	T	0

This can be re-written as

y_k	$Pr(Y = y_k)$
0	1/8
1	3/8
2	3/8
3	1/8

This can be re-written as

y_k	$Pr(Y = y_k)$
0	1/8
1	3/8
2	3/8
3	1/8

- This table represents a *binomial distribution* where we have $n = 3$ trials and the probability of success is $p = 0.5$.
- We can make similar tables for any value of n or p.

Parameters of the binomial distribution

- p = Probability of "success"
- n = # of "trials"

Parameters of the binomial distribution

- p = Probability of “success”
- n = # of “trials”

Mean and variance of the binomial

$$\mu = np$$

$$\sigma^2 = np(1 - p)$$

Example: Calculating the expected value of a binomial distribution

y_k	$Pr(Y = y_k)$
0	1/8
1	3/8
2	3/8
3	1/8

$$\mu = \sum_{k=1}^{K} y_k Pr(Y = y_k)$$

Example: Calculating the expected value of a binomial distribution

y_k	$Pr(Y = y_k)$
0	1/8
1	3/8
2	3/8
3	1/8

$$\mu = \sum_{k=1}^{K} y_k Pr(Y = y_k) = 0(1/8) + 1(3/8) + 2(3/8) + 3(1/8)$$

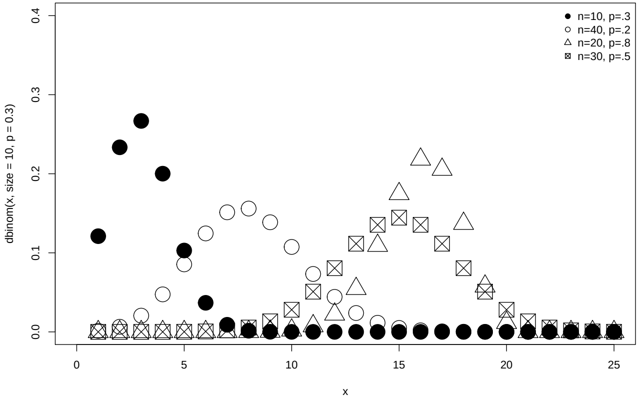
Example: Calculating the expected value of a binomial distribution

y_k	$Pr(Y = y_k)$
0	1/8
1	3/8
2	3/8
3	1/8

$$\mu = \sum_{k=1}^{K} y_k Pr(Y = y_k) = 0(1/8) + 1(3/8) + 2(3/8) + 3(1/8)$$

$$= 12/8 = 1.5 = np$$

```
x<-1:25
plot(x, dbinom(x, size=10, p=.3), cex=3, pch=19,
     ylim=c(0, .4), xlim=c(0, 25))
points(x, dbinom(x, size=40, p=.2), cex=3, pch=1)
points(x, dbinom(x, size=20, p=.8), cex=3, pch=2)
points(x, dbinom(x, size=30, p=.5), cex=3, pch=7)
legend("topright",
       c("n=10, p=.3", "n=40, p=.2", "n=20, p=.8", "n=30, p=.5"),
       pch=c(19,1,2,7), bty="n")
```



Continuous distributions

If a variable is continuous, we have a **probability density**

- $Pr(Y = a) = 0$

Continuous distributions

If a variable is continuous, we have a **probability density**

- $Pr(Y = a) = 0$
- $Pr(a \leq Y \leq b) = \int_b^a f(y)dy$

Continuous distributions

If a variable is continuous, we have a **probability density**

- $Pr(Y = a) = 0$
- $Pr(a \leq Y \leq b) = \int_b^a f(y)dy$
- $f(y) \geq 0 \; \forall \; y$

Continuous distributions

If a variable is continuous, we have a **probability density**

- $Pr(Y = a) = 0$
- $Pr(a \leq Y \leq b) = \int_b^a f(y)dy$
- $f(y) \geq 0 \ \forall \ y$
- $\int_Y f(y) = 1$

Continuous distributions

If a variable is continuous, we have a **probability density**

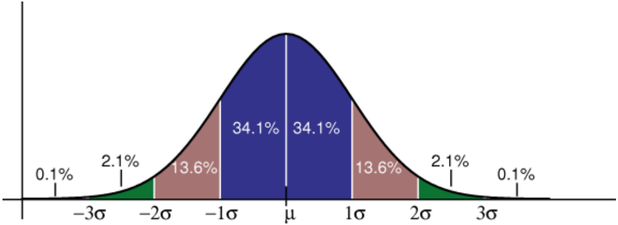
- $Pr(Y = a) = 0$
- $Pr(a \leq Y \leq b) = \int_b^a f(y)dy$
- $f(y) \geq 0 \; \forall \; y$
- $\int_Y f(y) = 1$
- $\mu = E(Y) = \int_Y y \; f(y)dy$

Continuous distributions

If a variable is continuous, we have a **probability density**

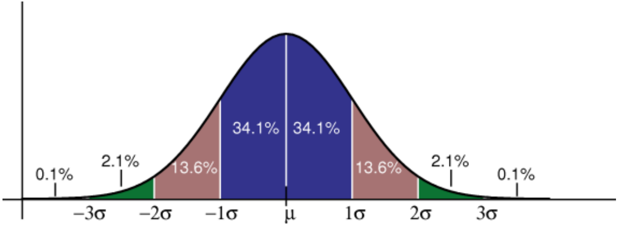
- $Pr(Y = a) = 0$
- $Pr(a \leq Y \leq b) = \int_b^a f(y)dy$
- $f(y) \geq 0 \; \forall \; y$
- $\int_Y f(y) = 1$
- $\mu = E(Y) = \int_Y y \; f(y)dy$
- $\sigma^2 = \ldots$

The normal distribution



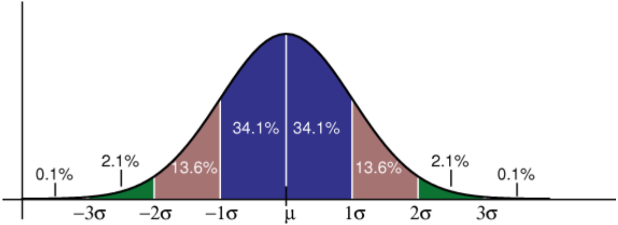
- Symmetric
- Bell shaped

The normal distribution



- Symmetric
- Bell shaped
- It's just a function

The normal distribution

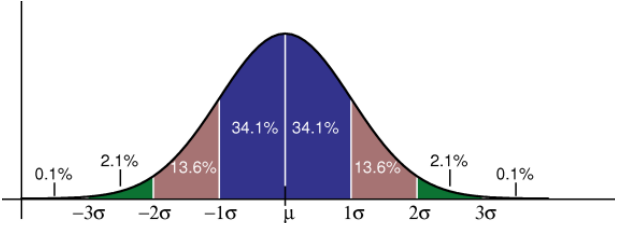


- Symmetric
- Bell shaped
- It's just a function

Normal distribution

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

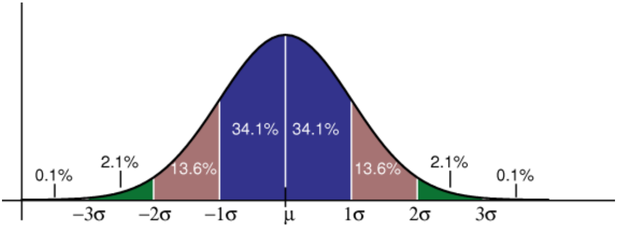
The normal distribution



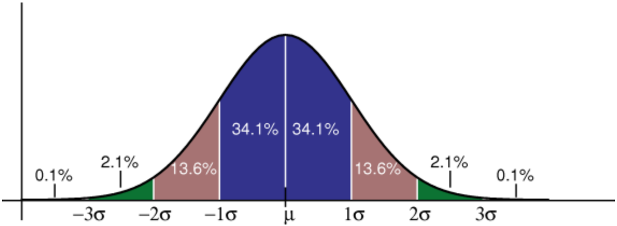
- Symmetric
- Bell shaped
- It's just a function

Normal distribution

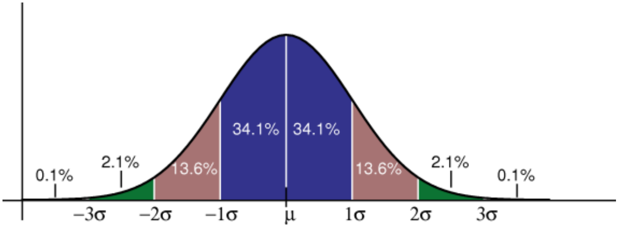
$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(y-\mu)^2} = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-1}{2\sigma^2}(y-\mu)^2\right)$$



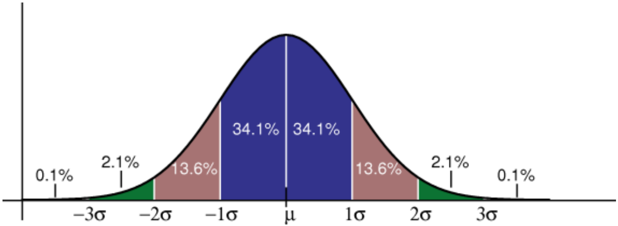
- Usually denoted $Y \sim N(\mu, \sigma^2)$



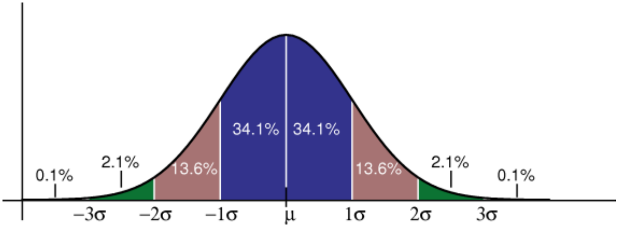
- Usually denoted $Y \sim N(\mu, \sigma^2)$
  - (note the two parameters)



- Usually denoted $Y \sim N(\mu, \sigma^2)$
  - (note the two parameters)
- This distribution is the basis of the "empirical rule."



- Usually denoted $Y \sim N(\mu, \sigma^2)$
  - (note the two parameters)
- This distribution is the basis of the "empirical rule."
- It is the one we teach you for reasons covered in the next several lectures (e.g., sampling distributions, Central Limit Theorem)



- Usually denoted $Y \sim N(\mu, \sigma^2)$
  - (note the two parameters)
- This distribution is the basis of the "empirical rule."
- It is the one we teach you for reasons covered in the next several lectures (e.g., sampling distributions, Central Limit Theorem) We cannot solve the integrals, but
  - tables will help you on exams
  - and `R` will help you on the homework.

T-Distribution

T-Distribution

- 1908 invented by William Gosset
- wanted to quickly test the quality of raw materials, testing small batches

Notes on the t-distribution

- It has “thicker”" tails than the normal distribution.
- Symmetric and bell-shaped

Notes on the t-distribution

- It has “thicker”" tails than the normal distribution.
- Symmetric and bell-shaped
- Dispersion depends on degrees of freedom, sometimes listed as *df* or *DOF*. If there is notation it is often v or ν.

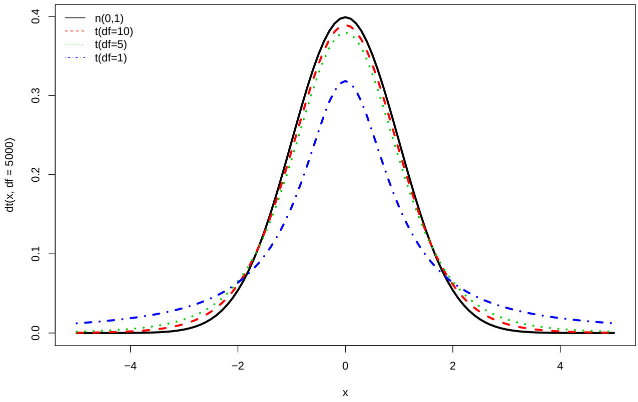
Notes on the t-distribution

- It has “thicker”" tails than the normal distribution.
- Symmetric and bell-shaped
- Dispersion depends on degrees of freedom, sometimes listed as *df* or *DOF*. If there is notation it is often *v* or ν.
- As $\nu \to \infty$ the t-distribution becomes essentially the normal distribution.

Notes on the t-distribution

- It has “thicker”" tails than the normal distribution.
- Symmetric and bell-shaped
- Dispersion depends on degrees of freedom, sometimes listed as *df* or *DOF*. If there is notation it is often *v* or ν.
- As $\nu \to \infty$ the t-distribution becomes essentially the normal distribution.
- NOTE: The use of the t-distribution is **not** related to the CLT. We are assuming the data is normally distributed.

```r
x<-seq(from=-5, to=5, by=.1)
plot(x, dt(x, df=5000), lwd=3, type="l", col=1, lty=1)
lines(x, dt(x, df=10), lwd=3, ylim=c(0, .4), col=2, lty=2)
lines(x, dt(x, df=5), lwd=3, ylim=c(0, .4), col=3, lty=3)
lines(x, dt(x, df=1), lwd=3, ylim=c(0, .4), col=4, lty=4)
legend("topleft",
       c("n(0,1)", "t(df=10)", "t(df=5)", "t(df=1)"),
       lty=c(1,2,3,4), col=c(1,2,3,4), bty="n")
```



Reviewing distributions

- What is the binomial distribution, and when would we use it?

Reviewing distributions

- What is the binomial distribution, and when would we use it?
- What is a probability distribution, and how is it different than a frequency distribution?

Reviewing distributions

- What is the binomial distribution, and when would we use it?
- What is a probability distribution, and how is it different than a frequency distribution?
- What are the important properties of a normal distribution?

Reviewing distributions

- What is the binomial distribution, and when would we use it?
- What is a probability distribution, and how is it different than a frequency distribution?
- What are the important properties of a normal distribution?
  - Symmetric
  - Bell shaped
  - Empirical rule

Reviewing distributions

- What is the binomial distribution, and when would we use it?
- What is a probability distribution, and how is it different than a frequency distribution?
- What are the important properties of a normal distribution?
  - Symmetric
  - Bell shaped
  - Empirical rule
- What is the t-distribution, and how is it different than a normal distribution?

Reviewing distributions

- What is the binomial distribution, and when would we use it?
- What is a probability distribution, and how is it different than a frequency distribution?
- What are the important properties of a normal distribution?
  - Symmetric
  - Bell shaped
  - Empirical rule
- What is the t-distribution, and how is it different than a normal distribution?
  - Thicker tails
  - Degrees of freedom instead of σ^2

When will we use these?

- Sometimes we will talk about how certain **statistics** are distributed

When will we use these?

- Sometimes we will talk about how certain **statistics** are distributed
  - Sampling distribution
- Sometimes we will talk about *assumptions* we make about how the **population** is distributed

When will we use these?

- Sometimes we will talk about how certain **statistics** are distributed
  - Sampling distribution
- Sometimes we will talk about *assumptions* we make about how the **population** is distributed
  - This can sometimes influence which sampling distribution we use

When will we use these?

- Sometimes we will talk about how certain **statistics** are distributed
  - Sampling distribution
- Sometimes we will talk about *assumptions* we make about how the **population** is distributed
  - This can sometimes influence which sampling distribution we use
- Try to keep it straight, although we will work on it

Class business

- Problem set 2 is now posted
- Review the online materials for lab (VERY IMPORTANT)
- The reading and online content for Wednesday is *absolutely essential*
- Questions or concerns?