Sampling Distributions

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Sampling Distributions

Introduction

- We learned a bit of probability theory.
- Now, let's try putting some of it to use.
- Before we can start using statistical analysis routinely to answer questions about data, there are a couple of intervening conceptual steps we need to take.

Introduction

- We learned a bit of probability theory.
- Now, let's try putting some of it to use.
- Before we can start using statistical analysis routinely to answer questions about data, there are a couple of intervening conceptual steps we need to take.

Introduction

- We learned a bit of probability theory.
- Now, let's try putting some of it to use.
- Before we can start using statistical analysis routinely to answer questions about data, there are a couple of intervening conceptual steps we need to take.

Introduction

- I want to begin with a question that is similar to one shown in your textbook.
- Suppose I throw two fair dice independently and completely at random, and then compute the *sum* of the two dice. This sum is the outcome.
- Is the probability distribution continuous or discrete?
- What is the probability distribution?

Introduction

- I want to begin with a question that is similar to one shown in your textbook.
- Suppose I throw two fair dice independently and completely at random, and then compute the *sum* of the two dice. This sum is the outcome.
- Is the probability distribution continuous or discrete?
- What is the probability distribution?

Introduction

- I want to begin with a question that is similar to one shown in your textbook.
- Suppose I throw two fair dice independently and completely at random, and then compute the *sum* of the two dice. This sum is the outcome.
- Is the probability distribution continuous or discrete?
- What is the probability distribution?

Introduction

- I want to begin with a question that is similar to one shown in your textbook.
- Suppose I throw two fair dice independently and completely at random, and then compute the *sum* of the two dice. This sum is the outcome.
- Is the probability distribution continuous or discrete?
- What is the probability distribution?

The Sum of Two Dice Throws

- Call the two dice X_1 and X_2.
- Considered individually, each has the same probability distribution.
- What is it?

The Sum of Two Dice Throws

- Call the two dice X_1 and X_2.
- Considered individually, each has the same probability distribution.
- What is it?

The Sum of Two Dice Throws

- Call the two dice X_1 and X_2.
- Considered individually, each has the same probability distribution.
- What is it?

The Sum of Two Dice Throws

- That's right — it is a discrete uniform, assigning probability 1/6 to each of the integers from 1 to 6, inclusive.
- But what happens if we throw two dice?
- Let's create a Venn diagram.
- This diagram will be a join events table, and I'll put the sum of the two dice throws in each interior cell.
- There are 36 interior cells. If the two dice throws are independent, what is the probability of each cell?

The Sum of Two Dice Throws

- That's right — it is a discrete uniform, assigning probability 1/6 to each of the integers from 1 to 6, inclusive.
- But what happens if we throw two dice?
- Let's create a Venn diagram.
- This diagram will be a join events table, and I'll put the sum of the two dice throws in each interior cell.
- There are 36 interior cells. If the two dice throws are independent, what is the probability of each cell?

The Sum of Two Dice Throws

- That's right — it is a discrete uniform, assigning probability 1/6 to each of the integers from 1 to 6, inclusive.
- But what happens if we throw two dice?
- Let's create a Venn diagram.
- This diagram will be a join events table, and I'll put the sum of the two dice throws in each interior cell.
- There are 36 interior cells. If the two dice throws are independent, what is the probability of each cell?

The Sum of Two Dice Throws

- That's right — it is a discrete uniform, assigning probability 1/6 to each of the integers from 1 to 6, inclusive.
- But what happens if we throw two dice?
- Let's create a Venn diagram.
- This diagram will be a join events table, and I'll put the sum of the two dice throws in each interior cell.
- There are 36 interior cells. If the two dice throws are independent, what is the probability of each cell?

The Sum of Two Dice Throws

- That's right — it is a discrete uniform, assigning probability 1/6 to each of the integers from 1 to 6, inclusive.
- But what happens if we throw two dice?
- Let's create a Venn diagram.
- This diagram will be a join events table, and I'll put the sum of the two dice throws in each interior cell.
- There are 36 interior cells. If the two dice throws are independent, what is the probability of each cell?

The Sum of Two Dice Throws

		X1					
		1	***2***	***3***	***4***	***5***	***6***
X2	***1***	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

- Note that, since the two dice are independent, the probability of a two is

$$\begin{aligned}
\Pr(X_1 + X_2 = 2) &= \Pr(X_1 = 1 \cap X_2 = 1) \\
&= \Pr(X_1 = 1) \times \Pr(X_2 = 1) \\
&= \frac{1}{6} \times \frac{1}{6} \\
&= \frac{1}{36}
\end{aligned}$$

The Sum of Two Dice Throws

		X1					
		1	***2***	***3***	***4***	***5***	***6***
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
X2	***4***	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

- Now let's compute the probability distribution of the sum of the dice throws.
- What is the probability that the sum is equal to 3?

The Sum of Two Dice Throws

X1

		1	2	3	4	5	6
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
X2	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

- Now let's compute the probability distribution of the sum of the dice throws.
- What is the probability that the sum is equal to 3?

The Sum of Two Dice Throws

X2 \ X1	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

- Notice that there are two ways one can obtain a 3.
- From the diagram, you should be able to see intuitively that the total probability of a sum of 3 is 2/36.
- Notice that the two "elementary events" are mutually exclusive.
- You cannot have a roll that is both $X_1 = 2 \cap X_2 = 1$ *and* $X_1 = 1 \cap X_2 = 2$.
- In set theory notation, we can make it look really complicated:

$$
\begin{aligned}
\Pr(\text{Sum} = 3) &= \Pr\left[(X_1 = 1 \cap X_2 = 2) \cup (X_1 = 2 \cap X_2 = 1)\right] \\
&= \Pr(X_1 = 1 \cap X_2 = 2) + \Pr(X_1 = 2 \cap X_2 = 1) \\
&= 1/36 + 1/36 \\
&= 1/18
\end{aligned}
$$

The Sum of Two Dice Throws

		X1					
		1	***2***	***3***	***4***	***5***	***6***
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
X2	***3***	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

- Notice that there are two ways one can obtain a 3.
- From the diagram, you should be able to see intuitively that the total probability of a sum of 3 is 2/36.
- Notice that the two "elementary events" are mutually exclusive.
- You cannot have a roll that is both $X_1 = 2 \cap X_2 = 1$ *and* $X_1 = 1 \cap X_2 = 2$.
- In set theory notation, we can make it look really complicated:

$$
\begin{aligned}
\Pr(\text{Sum} = 3) &= \Pr\left[(X_1 = 1 \cap X_2 = 2) \cup (X_1 = 2 \cap X_2 = 1)\right] \\
&= \Pr(X_1 = 1 \cap X_2 = 2) + \Pr(X_1 = 2 \cap X_2 = 1) \\
&= 1/36 + 1/36 \\
&= 1/18
\end{aligned}
$$

The Sum of Two Dice Throws

		X1					
		1	***2***	***3***	***4***	***5***	***6***
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
X2	***3***	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

- Notice that there are two ways one can obtain a 3.
- From the diagram, you should be able to see intuitively that the total probability of a sum of 3 is 2/36.
- Notice that the two "elementary events" are mutually exclusive.
- You cannot have a roll that is both $X_1 = 2 \cap X_2 = 1$ *and* $X_1 = 1 \cap X_2 = 2$.
- In set theory notation, we can make it look really complicated:

$$\begin{aligned}
\Pr(\text{Sum} = 3) &= \Pr\left[(X_1 = 1 \cap X_2 = 2) \cup (X_1 = 2 \cap X_2 = 1)\right] \\
&= \Pr(X_1 = 1 \cap X_2 = 2) + \Pr(X_1 = 2 \cap X_2 = 1) \\
&= 1/36 + 1/36 \\
&= 1/18
\end{aligned}$$

The Sum of Two Dice Throws

		X1					
		1	**2**	**3**	**4**	**5**	**6**
X2	**1**	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

- Notice that there are two ways one can obtain a 3.
- From the diagram, you should be able to see intuitively that the total probability of a sum of 3 is 2/36.
- Notice that the two "elementary events" are mutually exclusive.
- You cannot have a roll that is both $X_1 = 2 \cap X_2 = 1$ *and* $X_1 = 1 \cap X_2 = 2$.
- In set theory notation, we can make it look really complicated:

$$\begin{aligned}
\Pr(\text{Sum} = 3) &= \Pr\left[(X_1 = 1 \cap X_2 = 2) \cup (X_1 = 2 \cap X_2 = 1)\right] \\
&= \Pr(X_1 = 1 \cap X_2 = 2) + \Pr(X_1 = 2 \cap X_2 = 1) \\
&= 1/36 + 1/36 \\
&= 1/18
\end{aligned}$$

The Sum of Two Dice Throws

X1

X2		1	2	3	4	5	6
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

- Notice that there are two ways one can obtain a 3.
- From the diagram, you should be able to see intuitively that the total probability of a sum of 3 is 2/36.
- Notice that the two "elementary events" are mutually exclusive.
- You cannot have a roll that is both $X_1 = 2 \cap X_2 = 1$ *and* $X_1 = 1 \cap X_2 = 2$.
- In set theory notation, we can make it look really complicated:

$$\begin{aligned}
\Pr(\text{Sum} = 3) &= \Pr\left[(X_1 = 1 \cap X_2 = 2) \cup (X_1 = 2 \cap X_2 = 1)\right] \\
&= \Pr(X_1 = 1 \cap X_2 = 2) + \Pr(X_1 = 2 \cap X_2 = 1) \\
&= 1/36 + 1/36 \\
&= 1/18
\end{aligned}$$

The Sum of Two Dice Throws

		X1					
		1	***2***	***3***	***4***	***5***	***6***
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
X2	***3***	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

What is the probability of getting a sum of 4?

The Sum of Two Dice Throws

		X1					
		1	***2***	***3***	***4***	***5***	***6***
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
X2	***3***	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

What is the probability of getting a sum of 4?

The Sum of Two Dice Throws

- So now, on your own, you should be able to write the *entire probability distribution* for the Sum of Two Dice Throws.
- Write down all the probabilities for the various possible outcomes: 2,3,4,5,6,7,8,9,10,11,12 and put them in a table.

The Sum of Two Dice Throws

- So now, on your own, you should be able to write the *entire probability distribution* for the Sum of Two Dice Throws.
- Write down all the probabilities for the various possible outcomes: 2,3,4,5,6,7,8,9,10,11,12 and put them in a table.

The Sum of Two Dice Throws

Did it look like this?

Sum	*Pr(Sum)*
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

The Sum of Two Dice Throws

It DID? Great!

Sum	*Pr(Sum)*
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

The Sample Mean of Two Dice Throws

- Now I have a "big stretch" for you.
- Suppose that, instead of the *sum* of the two dice throws, I asked instead for the probability distribution of the *sample mean* M of the two dice throws.
- Can you give me *that* distribution?

The Sample Mean of Two Dice Throws

- Now I have a “big stretch” for you.
- Suppose that, instead of the *sum* of the two dice throws, I asked instead for the probability distribution of the *sample mean* M of the two dice throws.
- Can you give me *that* distribution?

The Sample Mean of Two Dice Throws

- Now I have a "big stretch" for you.
- Suppose that, instead of the *sum* of the two dice throws, I asked instead for the probability distribution of the *sample mean* M of the two dice throws.
- Can you give me *that* distribution?

The Sample Mean of Two Dice Throws

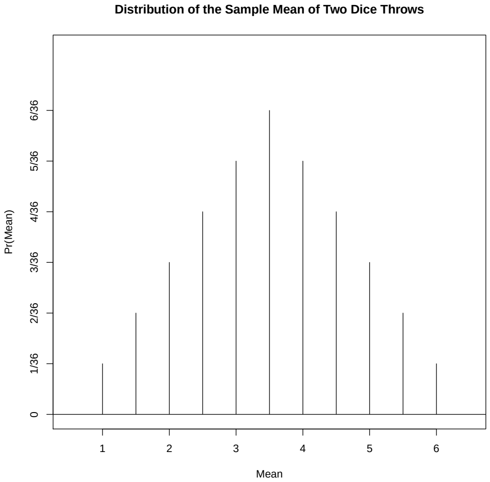
Of course you can!

Mean	*Pr(Mean)*
1	1/36
1.5	2/36
2	3/36
2.5	4/36
3	5/36
3.5	6/36
4	5/36
4.5	4/36
5	3/36
5.5	2/36
6	1/36

The Sample Mean of Two Dice Throws

Let's draw a graph of this distribution!

The Sample Mean of Two Dice Throws



Sampling Distributions

- The graph we just drew is our first example of a *sampling distribution*.
- What *is* a sampling distribution?

Sampling Distributions

- The graph we just drew is our first example of a *sampling distribution.*
- What *is* a sampling distribution?

Sampling Distributions

- Suppose you are interested in a *parameter* of a population, like the population mean.
- You take a sample, and estimate the parameter with a *statistic*, such as the sample mean.
- Now you ask yourself the question: Suppose I were to repeat this operation over and over and over. Take a sample, compute the mean. Take a sample, compute the mean, take a sample, compute the mean, each time keeping track of the values of those sample means.
- If I drew the distribution of those values, in the long run that distribution would converge to the *sampling distribution of the sample mean.*

Sampling Distributions

- Suppose you are interested in a *parameter* of a population, like the population mean.
- You take a sample, and estimate the parameter with a *statistic*, such as the sample mean.
- Now you ask yourself the question: Suppose I were to repeat this operation over and over and over. Take a sample, compute the mean. Take a sample, compute the mean, take a sample, compute the mean, each time keeping track of the values of those sample means.
- If I drew the distribution of those values, in the long run that distribution would converge to the *sampling distribution of the sample mean.*

Sampling Distributions

- Suppose you are interested in a *parameter* of a population, like the population mean.
- You take a sample, and estimate the parameter with a *statistic*, such as the sample mean.
- Now you ask yourself the question: Suppose I were to repeat this operation over and over and over. Take a sample, compute the mean. Take a sample, compute the mean, take a sample, compute the mean, each time keeping track of the values of those sample means.
- If I drew the distribution of those values, in the long run that distribution would converge to the *sampling distribution of the sample mean.*

Sampling Distributions

- Suppose you are interested in a *parameter* of a population, like the population mean.
- You take a sample, and estimate the parameter with a *statistic*, such as the sample mean.
- Now you ask yourself the question: Suppose I were to repeat this operation over and over and over. Take a sample, compute the mean. Take a sample, compute the mean, take a sample, compute the mean, each time keeping track of the values of those sample means.
- If I drew the distribution of those values, in the long run that distribution would converge to the *sampling distribution of the sample mean.*

Sampling Distributions

Let's try to bring this to life with some R code!

Sampling Distributions

- R has a really neat function called `replicate` that automatically does something over and over, as many times as you want (and have time to wait for).
- Let's see how it works.

Sampling Distributions

- Suppose we were not smart enough to derive the sampling distribution of sample means for two throws of a fair die.
- Instead, we use R to simulate this sampling distribution.
- To do that, we first simulate a single random die throw. Notice I set the seed of the random number generator first to make sure I can replicate my exercise.

```
> set.seed(12345)
> sample(1:6,1)
[1] 5
```

- I threw the die and got a 5.
- Now I'll throw two dice.

```
> sample(1:6,2,replace=TRUE)
[1] 6 5
```

- I got a 6 and a 5.

Sampling Distributions

- Let me restart the random number generator and throw two dice.

```
> set.seed(12345)
> sample(1:6,2,replace=TRUE)
[1] 5 6
```

- I got a 5 and a 6.
- Now, I could have asked R to throw the two dice and sum them. Here's how I could do that.

```
> set.seed(12345)
> sum(sample(1:6,2,replace=TRUE))
[1] 11
```

- But I also ask R to throw the two dice and compute their mean.

```
> set.seed(12345)
> mean(sample(1:6,2,replace=TRUE))
[1] 5.5
```

Sampling Distributions

- Let me restart the random number generator and use the **`replicate`** command to ask R to throw two dice 25 times, each time computing and reporting the mean.
- Here we go.

```
> set.seed(12345)
> replicate(25,mean(sample(1:6,2,replace=TRUE)))
 [1] 5.5 5.5 2.0 3.0 5.5 1.0 3.0 3.0 3.0 4.0 2.5 5.5 3.
[20] 2.5 4.0 5.5 2.0 1.0 2.5
```

Sampling Distributions

- Of course, we could have asked for 100 replications.

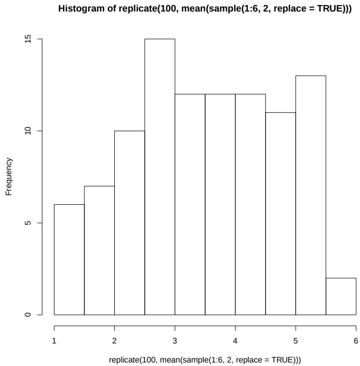
```
> set.seed(12345)
> replicate(100,mean(sample(1:6,2,replace=TRUE)))
```

```
  [1] 5.5 5.5 2.0 3.0 5.5 1.0 3.0 3.0 3.0 4.0 2.5 5.5 3.5 4.5 2.5 3.0 3.5 3.0
 [19] 6.0 2.5 4.0 5.5 2.0 1.0 2.5 5.5 2.0 4.0 3.0 2.5 3.5 5.5 4.0 3.5 5.0 4.5
 [37] 3.0 2.5 5.0 2.5 5.0 1.0 1.5 4.5 3.0 5.5 2.5 4.5 5.0 2.0 3.0 5.0 5.0 5.0
 [55] 2.0 3.5 4.0 5.5 5.0 1.5 4.0 3.0 5.0 4.0 3.5 3.0 4.5 6.0 3.5 3.0 4.5 3.0
 [73] 3.5 4.5 4.5 4.5 4.0 3.5 4.0 5.0 4.5 5.5 3.0 4.0 4.5 4.0 3.5 4.0 5.5 5.0
 [91] 2.5 2.0 5.5 4.5 1.0 5.5 2.0 3.5 3.5 2.5
```

Sampling Distributions

- Rather than listing them all out, we could call for a histogram. Unfortunately, R's histogram routine is not very good for graphing discrete distributions.

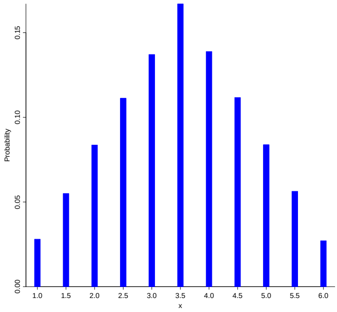
```
> set.seed(12345)
> hist(replicate(100,mean(sample(1:6,2,replace=TRUE))))
```



Sampling Distributions

- We're going to need a bigger number of replications and a better method of graphing the discrete histogram.

```
> set.seed(12345)
> data <- replicate(100000,mean(sample(1:6,2,replace=TRUE)))
> x <- 2:12 / 2
> px <- rep(0,11)
> for(i in 1:11) px[i] <- mean(data == x[i])
> discrete.histogram(x,px)
```



Sampling Distributions

- So, we already have two distinct methods to learn about sampling distributions:
  1. A derivation using probability theory, and
  2. A brute force approach using *Monte Carlo simulation*.

Facts about Sampling Distributions

- Sampling distributions often do not have the same shape as the population from which the sample was taken.

Example (Distribution of the Mean)

We just saw that the sampling distribution of the mean of two dice throws has a shape that is triangular, and more reminiscent of the shape of a normal distribution than the shape of the population from which the samples of size 2 were taken. That distribution, the distribution of the population, was discrete uniform, ranging over the integers from 1 to 6.

Facts about Sampling Distributions

- The shape of a sampling distribution may change as a function of the sample size n.
- Even if nothing else changes, the variability of the sampling distribution decreases as a function of n for all statistics in common use.
- Exact sampling distributions are difficult to get. Often, we have to rely on approximations and *asymptotic results*
- Perhaps the most famous result in sampling distribution theory is the *Central Limit Theorem*.

Facts about Sampling Distributions
The Central Limit Theorem

Example (The Central Limit Theorem)

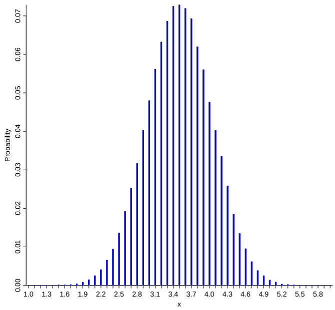
Suppose you take independent samples of size n from a distribution that has a finite variance σ^2 and mean μ. Then as n becomes large, the shape of the distribution of sample means converges to the standard normal distribution with a mean of μ and a standard deviation $\sigma/\sqrt{n}$, regardless of the shape of the population distribution. However, if the population distribution is normal in shape, then the sampling distribution will be normal, regardless of the sample size.

Facts about Sampling Distributions
The Central Limit Theorem

- The "Central Limit Theorem Effect" can be seen operating in our dice throw example.
- The distribution of the individual dice is flat, but even with an n of 2, the distribution of the sample mean is already moving toward a normal distribution in shape.
- Suppose we threw ten dice simultaneously. What would the distribution look like? Modifying the code I used before, I take a look using the brute force approach.

Facts about Sampling Distributions
The Central Limit Theorem

```
> data <- replicate(100000,mean(sample(1:6,10,replace=TRUE)))
> x <- 10:60 / 10
> px <- rep(0,51)
> for(i in 1:51) px[i] <- mean(data == x[i])
> discrete.histogram(x,px)
```



Facts about Sampling Distributions

The Central Limit Theorem

- The Central Limit Theorem effect explains why many quantities are, at least through the middle 4 standard deviations, well approximated by the normal distribution.
- We can also use the result, and our knowledge of normal curves, to do some straightforward calculations.

Facts about Sampling Distributions
The Central Limit Theorem

- Suppose you are doing research on the average height of Vanderbilt male students, and you know that the population standard deviation is around 2.5. You take a random sample of size $n = 25$. If the population mean is 70, what proportion of the time will your sample mean be within one inch of the correct value?
- Note that this is a long-run probability we are calculating. Anything might happen with your particular sample.

Facts about Sampling Distributions

The Central Limit Theorem

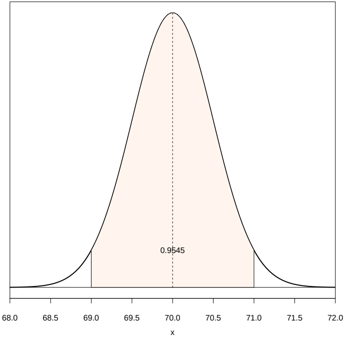
- We want to know what proportion of the time the sample mean is between 69 and 71. The Central Limit Theorem result tells us that the distribution of the sample mean will be approximately normal, with a mean of 70, and a standard deviation of

$$\sigma_M = \sigma/\sqrt{n} = 2.5/\sqrt{25} = 0.5.$$

Facts about Sampling Distributions
The Central Limit Theorem

- In R, we can use our normal curve drawing and coloring routine to display the result.

```
> nc(70,0.5)
> cn(69,71,70,0.5,col=colors()[579])
```



Facts about Sampling Distributions
The Central Limit Theorem

- This result implies the following.
- A bit more than 95% of the time, in the long run, when $\sigma = 2.5$ and $n = 25$, the sample mean will be within 1 point of 70 if $\mu = 70$.
- But suppose you don't know μ, which will be the case most of the time when you run an experiment.
- What can you say then?

Facts about Sampling Distributions

The Signal and the Noise

- The sampling distribution of M shows how it behaves over repeated samples.
- We can view μ, the quantity that M is estimating, as the "signal" in our sample, and the variability around μ as the "noise level."
- All other things being equal, how does n affect the noise level?

Facts about Sampling Distributions

The Signal and the Noise

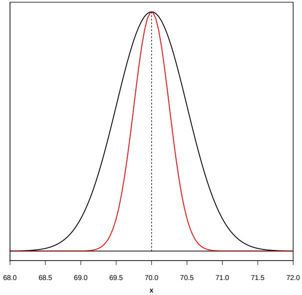
- Suppose I quadrupled n and made it 100. What would happen?

Facts about Sampling Distributions

The Signal and the Noise

In an important sense, by quadrupling n we made M twice as precise in its long run estimation of μ.

```
> nc(70,.5) # with n = 25
> # Now draw curve with n=100 in red
> draw.normal(68,72,70,.25,col="red",new=TRUE)
```



Facts about Sampling Distributions

The Signal and the Noise

```
> nc(70,2.5/sqrt(100))
> cn(69.5,70.5,70,2.5/sqrt(100),col=colors()[579])
```