Student t-Test

ECE 341 Random Processes

Jake Glower - Lecture #25

Please visit Bison Academy for corresponding
lecture notes, homework sets, and solutions

Calculating Probabilities

In this course, we've covered several ways to calculate probabilities.

For example, let

$$Y = 4d4 + 3d6 + 2d8$$

Determine

- $p(y = 35)$
- $p(y > 35)$
- 90% confidence interval for y

Option 1: Monte-Carlo

Roll the dice 1 million times

Count the results

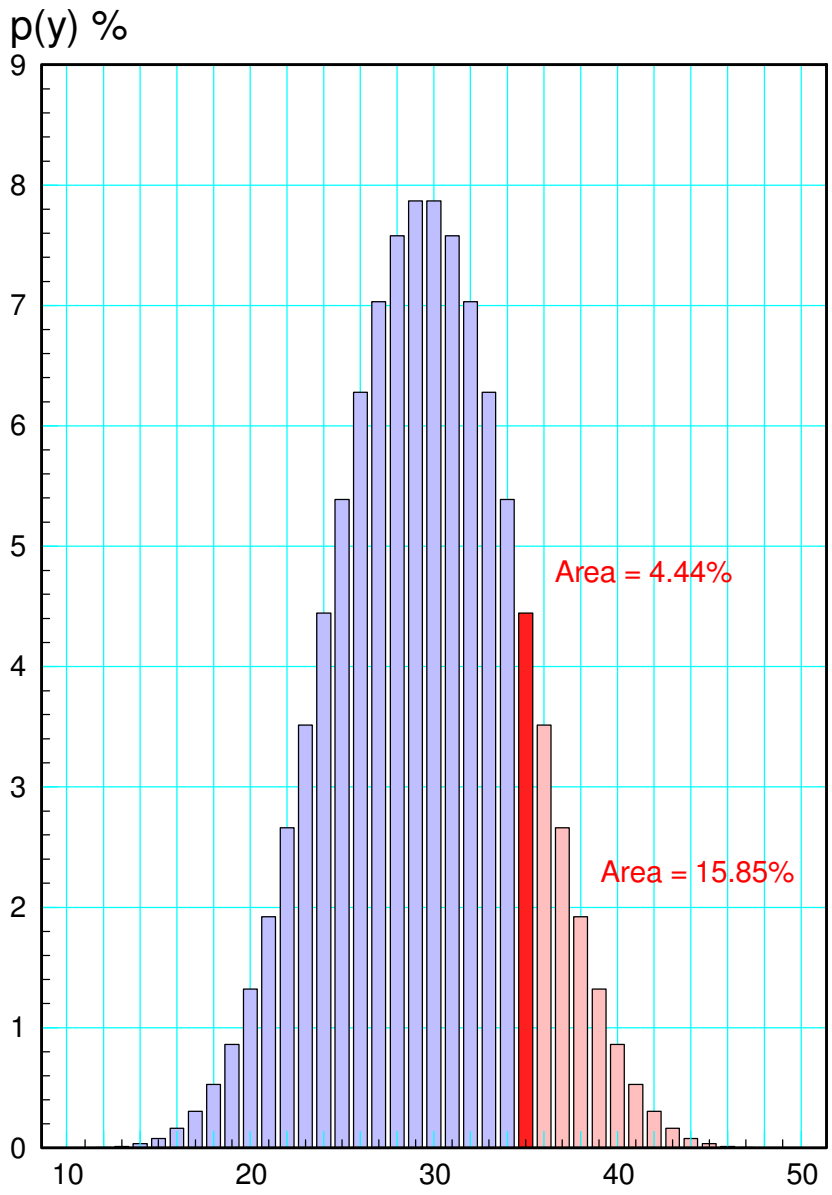
- pdf shown to the right

Results:

- p(y = 35) = 4.44%
- p(y > 35) = 15.85%
- p(21.5 < y < 36.5) = 90%

note:

- This took 1,000,000 rolls
- At $1/roll, this costs $1,000,000



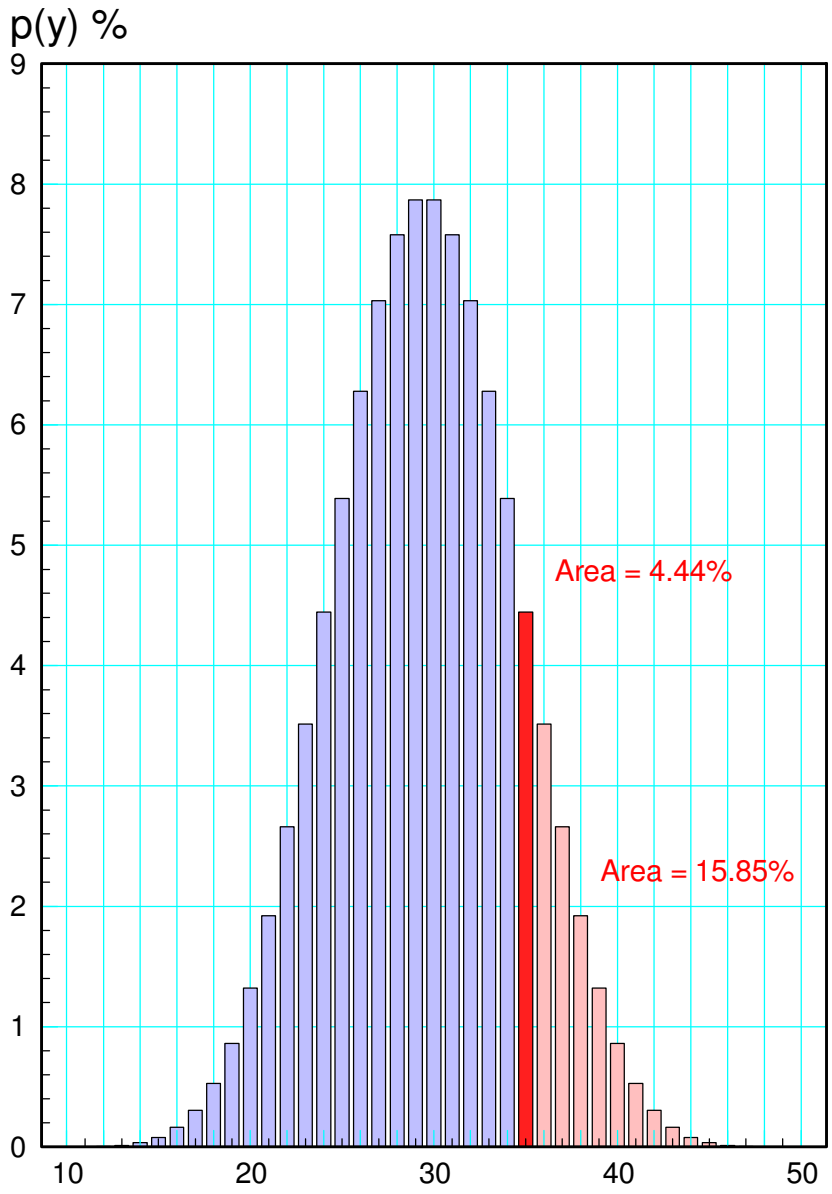
Option 2: Calculate the odds

Several ways to do this

- Enumeration
- Convolution
- Combinatorics

These all give exact answers

You have to know the pdf for these to work, however



Option 3: Normal Approximation

We *know* the mean and variance for a single die

We can *calculate* the mean and variance for y

	d4	d6	d8	4d4 + 3d6 + 2d8
mean	2.5	3.5	4.5	29.5
variance	1.25	2.9167	5.25	24.25

Normal Approximation (cont'd)

Once you know the mean and variance, you can calculate the odds

- Determine the z-score
- From a t-Table convert to a probability

$p(y < 34.5) = 0.8450$

$p(y < 35.5) = 0.9995$

so

$p(34.5 < y < 35.5) = 0.0434$

90% confidence interval

- $21.4 < y < 37.6$

```
>> z1 = (34.5-29.5)/sqrt(24.25)
z1 =    1.0153

>> p1 = (erf(z1/sqrt(2)) + 1)/2
p1 =    0.8450

>> z2 = (35.5-29.5)/sqrt(24.25)
z2 =    1.2184

>> p2 = (erf(z2/sqrt(2)) + 1)/2
p2 =    0.8885

>> p = p2 - p1
p =    0.0434

>> p3 = (erf(z3/sqrt(2)) + 1)/2
p3 =    0.1115

>> 29.5 + 1.64485 * sqrt(24.25)
ans =   37.5999

>> 29.5 - 1.64485 * sqrt(24.25)
ans =   21.4001
```

Normal Approximation: Results

With a Normal approximation, you get

- Almost the same results
- With zero die rolls

	Monte-Carlo	Normal Approx
p(y = 35)	4.43%	4.34%
p(y > 35)	11.39%	11.15%
90% confidnce interval	[22, 37]	(21.4, 37.6)
# rolls	1,000,000	0

Problem

What if

- You don't know the pdf?
- You don't know the mean?
- You don't know the variance?

But you can collect measurements...

Solution (Option 4): Student t Distribution

Assume a Normal distibution

- Usually the case

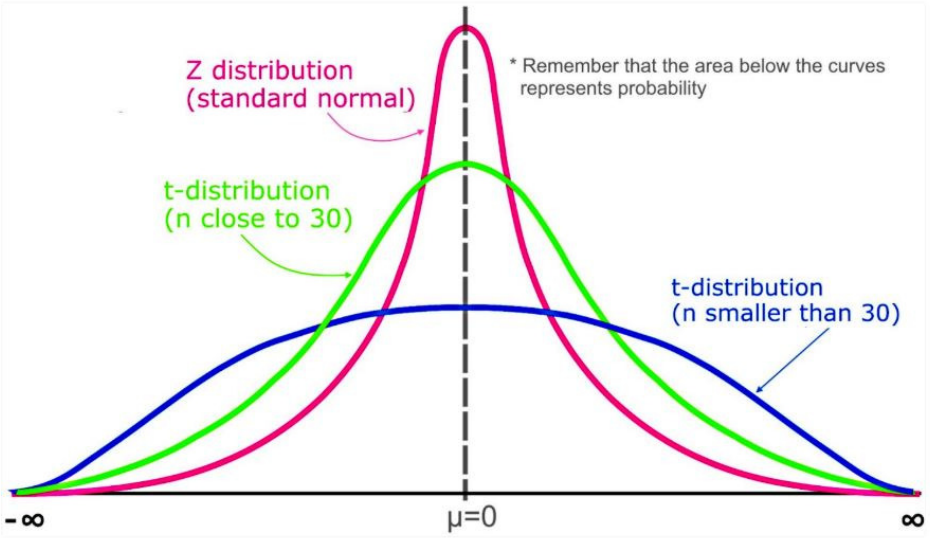
Collect n samples from the population

From these samples estimate

- The mean
- The variance

The result is a *Student t Distribution*

- Very similar to a Normal distribution
- Takes sample size into account



Normal vs. Student t Distribution

A Normal distribution is defined by two parameters

$$\mu = \frac{1}{n}\sum x_i \qquad \textit{mean}$$

$$\sigma^2 = \frac{1}{n}\sum (x_i - \mu)^2 \qquad \textit{variance}$$

Probabilities are computed using a z-score

$$z = \left(\frac{x-\mu}{\sigma}\right)$$

A Student t-Distribution is defined by three parameters

$$\bar{x} = \frac{1}{n}\sum x_i \qquad \textit{mean}$$

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 \qquad \textit{variance}$$

$$dof = n - 1 \qquad \textit{degrees of freedom}$$

Probabilities are computed using a t-score

$$t = \left(\frac{x-\bar{x}}{s}\right)$$

t-Table

Left side

- degrees of freedom
- sample size minus 1

Top

- Area of the tail

Middle

- t-score
- similar to a z-score

Note

- t-score is equal to the z-score for infinite dof

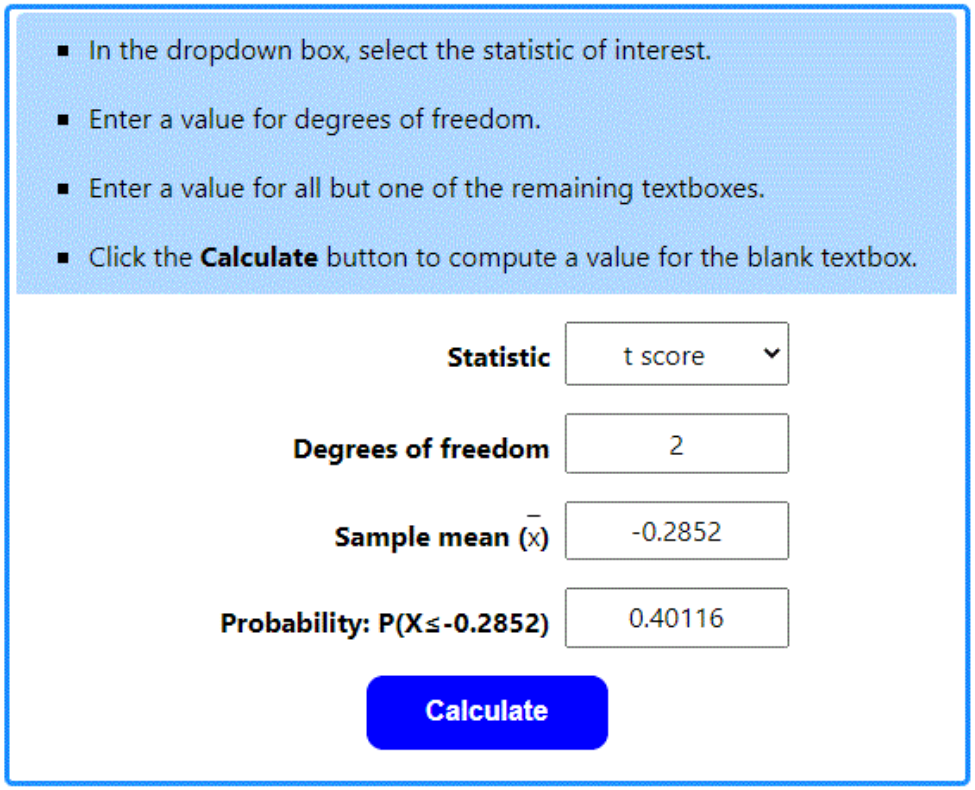
df \ p	0.001	0.0025	0.005	0.01000	0.025	0.05000
1	636.61900	318.30900	63.65670	31.82050	12.70620	6.31380
2	31.59910	22.32710	9.92480	6.96460	4.30270	2.92000
3	12.92400	10.21450	5.84090	4.54070	3.18240	2.35340
4	8.61030	7.17320	4.60410	3.74690	2.77640	2.13180
5	6.86880	5.89340	4.03210	3.36490	2.57060	2.01500
6	5.95880	5.20760	3.70740	3.14270	2.44690	1.94320
7	5.40790	4.78530	3.49950	2.99800	2.36460	1.89460
8	5.04130	4.50080	3.35540	2.89650	2.30600	1.85950
9	4.78090	4.29680	3.24980	2.82140	2.26220	1.83310
10	4.58690	4.14370	3.16930	2.76380	2.22810	1.81250
100	3.39050	3.17370	2.62590	2.36420	1.98400	1.66020

t-Table (StatTrek)

You can also use StatTrek for t Tables

- Input degrees of freedom
- Input eith the t-score or
- The probability
- Press Calculate

StatTrek computes and displays the remaining term



Note on t Tables

A sample size of 1 is meaningless

- Zero degrees of freedom
- You can't estimate two numbers (mean and variance) from a single sample

A sample size of 2 works

- t-scores are very large
- Reflects uncertainty with only 2 measurements

More samples helps

- t-score gets smaller
- Diminishing returns

df \ p	0.001	0.0025	0.005	0.01000	0.025	0.05000
1	636.61900	318.30900	63.65670	31.82050	12.70620	6.31380
2	31.59910	22.32710	9.92480	6.96460	4.30270	2.92000
3	12.92400	10.21450	5.84090	4.54070	3.18240	2.35340
4	8.61030	7.17320	4.60410	3.74690	2.77640	2.13180
5	6.86880	5.89340	4.03210	3.36490	2.57060	2.01500
6	5.95880	5.20760	3.70740	3.14270	2.44690	1.94320
7	5.40790	4.78530	3.49950	2.99800	2.36460	1.89460
8	5.04130	4.50080	3.35540	2.89650	2.30600	1.85950
9	4.78090	4.29680	3.24980	2.82140	2.26220	1.83310
10	4.58690	4.14370	3.16930	2.76380	2.22810	1.81250
100	3.39050	3.17370	2.62590	2.36420	1.98400	1.66020

Dice: Sample Size = 1

Let's go back to rolling dice

$$Y = 4d4 + 3d6 + 2d8$$

Determine the

- Probability that y = 35,
- Probability that y > 35, and
- 90% confidence interval

using a single die roll

```
n = 1;
Roll = zeros(n,1);
for i=1:1
   d4 = ceil( 4*rand(1,4) );
   d6 = ceil( 6*rand(1,3) );
   d8 = ceil( 8*rand(1,2) );
   Y = sum(d4) + sum(d6) + sum(d8);
   Roll(i) = Y;
   end

Roll = 30
```

t-Test Calculations

To use a t-Test, you first compute the mean and variance:

$$\bar{x} = \frac{1}{n}\sum x_i = 30$$

$$s^2 = \frac{1}{n-1}\sum (x - \bar{x})^2 = \frac{0}{0}$$

With a sample size of one, the varianec is undefined. This tells you:

You cannot determine probabilities using a single measurement.

You need at least two measurements to do any analysis.

t Test with Sample Size of 3

Now you *can* get results

```
Roll = 30    40    28

x = 32.6667

s = 6.4291

n = 3
```

```
n = 3;
Roll = zeros(n,1);
for i=1:1
   d4 = ceil( 4*rand(1,4) );
   d6 = ceil( 6*rand(1,3) );
   d8 = ceil( 8*rand(1,2) );
   Y = sum(d4) + sum(d6) + sum(d8);
   Roll(i) = Y;
   end

x = mean(Roll)
s = std(Roll)
n = length(Roll)
```

t Test: p(y = 35)

Calculate the t-scores

Use a t Table to convert to probabilities

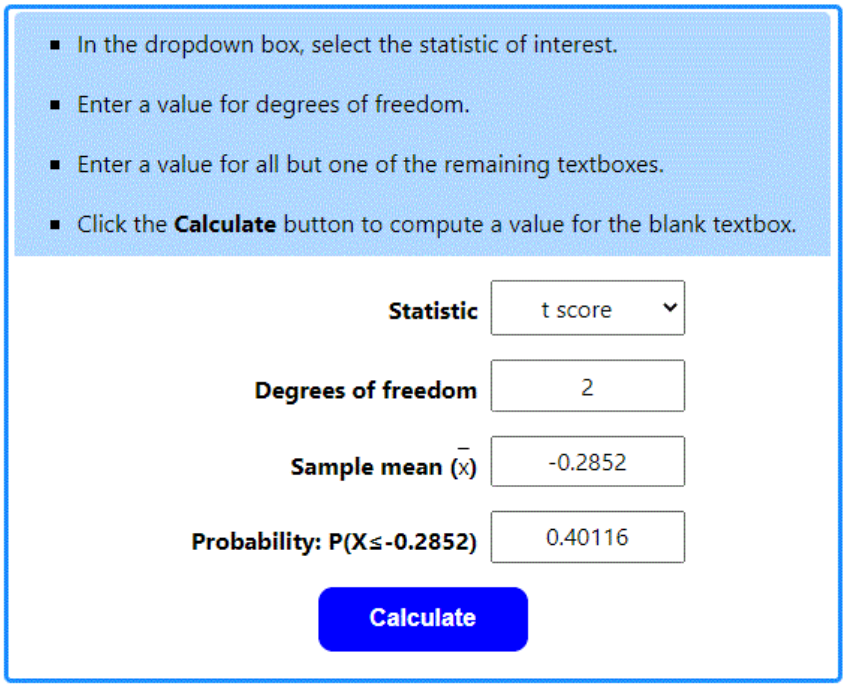
```
>> t1 = (34.5 - x) / s
t1 =    0.2852

>> t2 = (35.5 - x) / s
t2 =    0.4407
```

From StatTrek, with 2 dof

- Degrees of Freedom = 2 (Sample Size - 1)
- p1 = 40.116%
- p2 = 35.124%

Difference = 4.992%

- vs. 4.4445% (exact)



t-test: p(y>35)

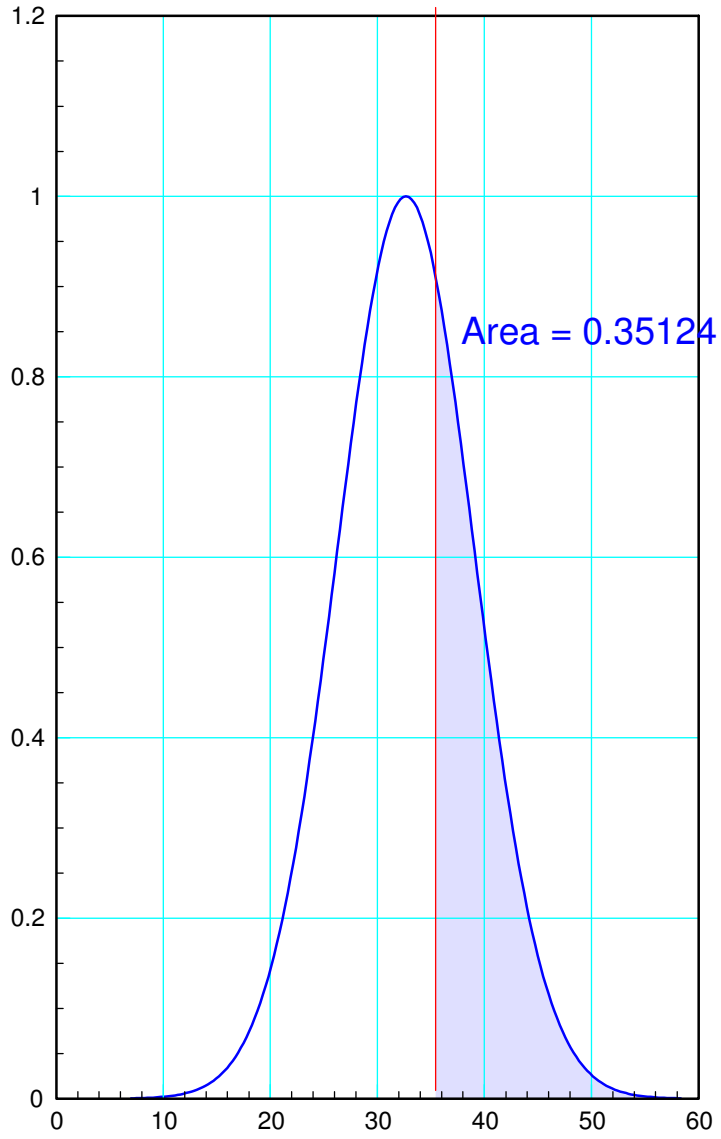
- p(y > 35.5)

Calculate the distance to the mean

```
>> t2 = (35.5 - x) / s
t2 =    0.4407
```

Convert this to a probability using a Student-t table

- StatTrek also works
- Degrees of Freedom = 2 (Sample Size - 1)
- p2 = 35.124%
- exact = 15.8524%



t Test: 90% Confidence Interval

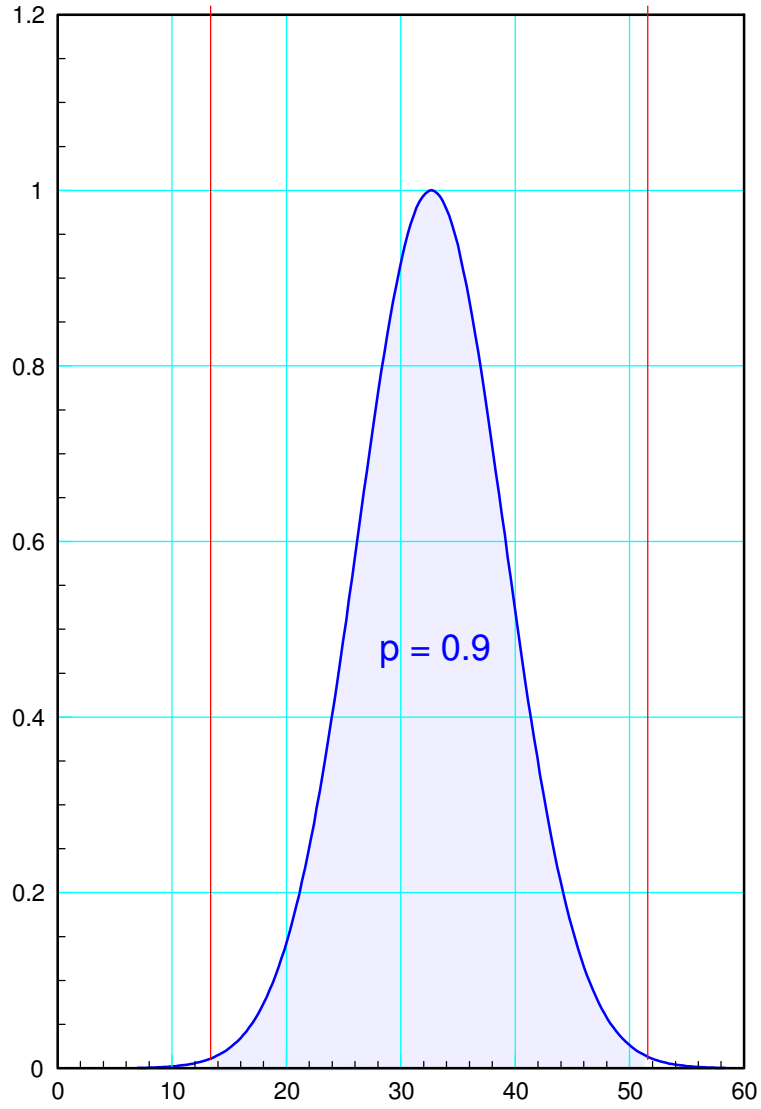
Determine the t-score for

- 2 degrees of freedom
- 5% tails
- t = 2.9200

Go left and right 2.92 standard deviations

Result

- 13.89 < y < 51.40 (p = 90%)
- 21.5 < y < 38.5 (from enumeration)



Results

Similar results to Monte Carlo

- 3 rolls vs. 1,000,000

	Monte-Carlo	Normal Approx	t-Test
p(y = 35)	4.43%	4.34%	4.992%
p(y > 35)	11.39%	11.15%	35.12%
90% confidnce interval	22 < y < 37	21.4 < y < 37.6	13.89 < y < 41.54
# rolls	1,000,000	0	3

Results get better with larger sample size

	# Rolls	p(y = 35)	p(y >= 35)	90% conf interval
Enumeration (exact)	3,538,944	4.4445%	15.8524%	(21.5, 38.5)
Monte-Carlo	100,000	4.444%	15.859%	(21.5, 38.5)
Normal Approx	0	4.344%	15.498%	(21.4, 37.6)
t-Test	3	4.992%	35.12%	(13.9, 41.5)
t-Test	10	3.75%	18.057%	(18.2, 39.5)
t-Test	30	4.27%	14.17%	(22.1, 37.2)

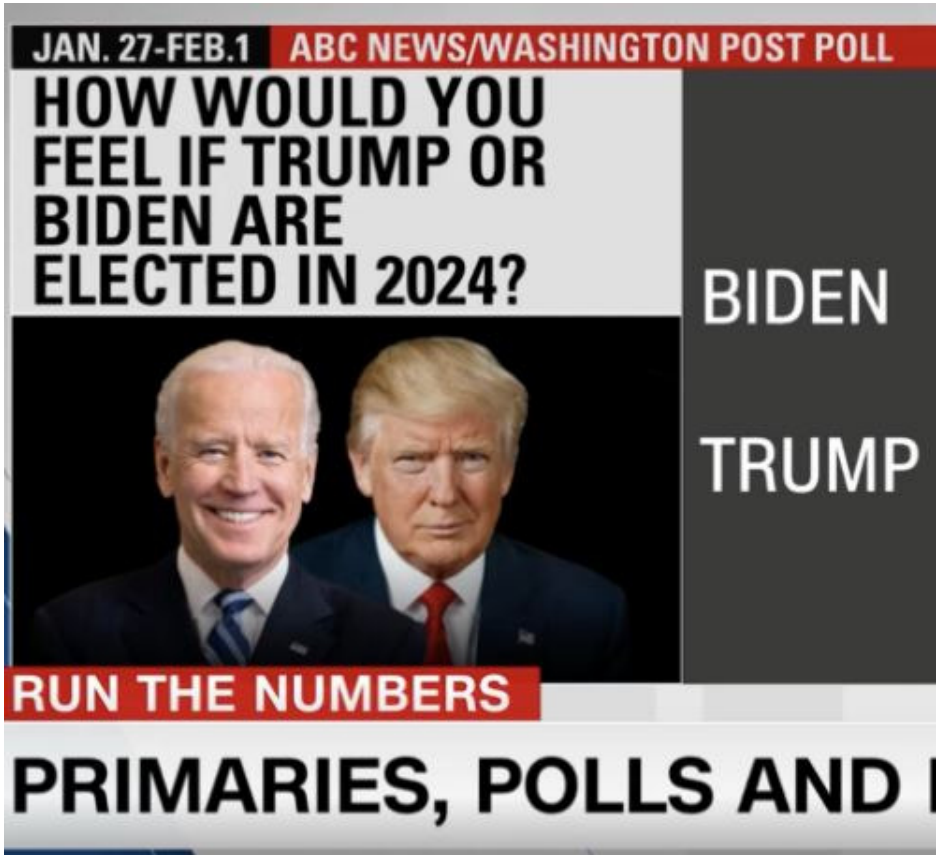
Sample Size & t-Tests

You don't need a huge sample size

- 2023 ABC News poll:
- 1006 adults
  - *National poll for 350 million people*

CNN Poll

- October 23-28, 2024
- 726 voters in Michigan polled
  - *5,662,504 votes total*
- 819 voters in Pennsylvania polled
  - *7,034,206 votes total*
- 736 voters in Wisconsin polled
  - *3,415,213 votes total*



But - Data Needs to be Unbiased.

- Bad data can give bad results
- You might be measuring something else

Example: Great Depression

- Telephone poll to see how many are unemployed
- Result: Everyone had a job

Problem:

- If you don't have a job, you reduce expenses
- Phones are one of the first things to go

Actual Survey Result:

- People who can afford a telephones have jobs

Example 2: 2016 Michigan Primary

- Hillary vs. Bernie

Telephone polls predicted a Clinton victory

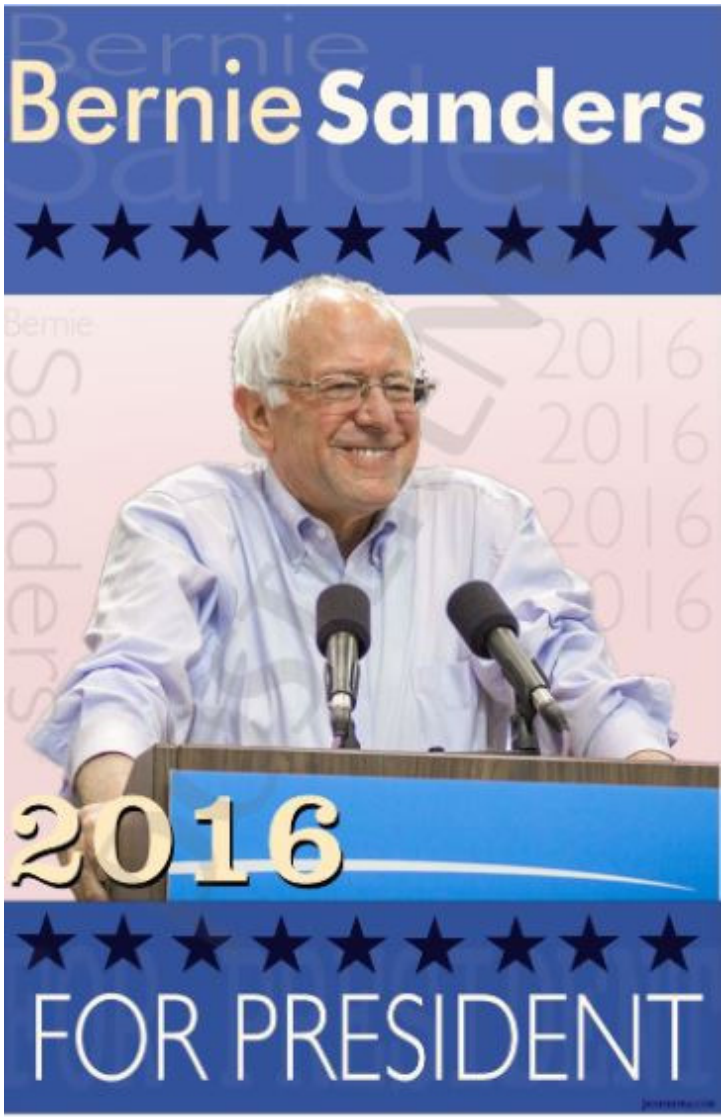
- Bernie Sanders actually won
- Complete surprise to everyone

Problem:

- Older voters have land lines
- Younger voters have cell phones
- Polls only used land lines

Actual Survey Results:

- Older voters support Clinton



How to Get Unbiased Data?

- Actually a difficult problem
- Why pollsters are highly skilled professionals

Example: ABC News / Washington Post Poll

- Sept 15-20, 2023
- English and Spanish
- 1006 Adults
- 25-25-42% Democrat-Republican-Independent

Polling isn't easy

- Great pains to make sure sample is random
- Non-responses causes a bias in the data
- Results are weighted to account for this bias

Accurate polling is a highly developed, highly skilled process

Populations vs. Individuals

There's a *slight* difference when asking questions about individuals vs. populations

Individuals: Matlab

$$s^2 = \left(\frac{1}{n-1}\right)\sum (x_i - \bar{x})^2$$

```
v = var(Data)
```

Populations:

$$s^2 = \left(\frac{1}{n}\right)\left(\frac{1}{n-1}\right)\sum (x_i - \bar{x})^2$$

```
v = var(Data) / n
```

If you want to know the variation of a single roll, use the formula for individuals

If you want to know the value of the population's mean, use the population formula

- You know more about populations than individuals
- As sample size goes to infinity, you know the population's mean *exactly*
- Individuals still have variability

What is the mean of y?

- Y = 4d4 + 3d6 + 2d8
- Population question
- Divide the variance by n

Sample Size	t-score (5% tails)	Population's Mean
1	-	undefined
3	2.9200	23.4718 < mean < 44.5282
10	1.83110	25.4239 < mean < 31.9761
100	1.66039	28.1737 < mean < 29.8263
1,000	1.64838	29.2831 < mean < 29.8029
infinite	1.64485	29.5000 < mean < 29.5000 *exact*

When do you divide the variance by sample size?

- Depends upon what question you're asking
- Individual vs. Population

Individual:

- *How many times will I get a full-house in 100,000 hands of poker?*
- Variation remains as sample size goes to infinity
- Sample variance convergers to the population's variance

Population:

- *What are the odds of being dealt a full house in poker?*
- You're looking for a number
- Confidence interval goes to zero as sample size goes to infinity

Summary

If you want to know what's coming off the assembly line, you need data.

If you measure everything (enumeration),

- You know what you're producing, but
- You have no product (nothing is new)

With a Monte-Carlo simulation,

- You get good results,
- But you need a large sample size (can be expensive)

If you know the mean and standard deviation...

- Use a Normal approximation
- Requires zero measurements

If you don't know the mean and standard deviation...

- Use a Student-t Test
- Requires a small sample size ($n \geq 2$)
- More data helps, but you don't need a huge amount of data