Student t-Test

ECE 341 Random Processes

Jake Glower - Lecture #23a

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

Calculating Probabilities

If all outcomes are equally likely, you can calculate probabilities using enumeration

- Convolution is one way of enumeration all possibilities
- If you measure everything you produce, you know what your product is
- You also are broke since you have no product to sell

If you know the mean and standard deviation,

- Assume a normal distribution (Central Limit Theorem)
- Calculate z-scores & probabilities
- Requires no measurements

If you do not know the mean and standard deviation,

- Assume a normal distribution (Central Limit Theorem)
- Collect data and estimate the mean and standard deviation (small sample size)
- The resulting distribution is a Student-t distribution
- Similar to a normal distribution, but takes sample size into account

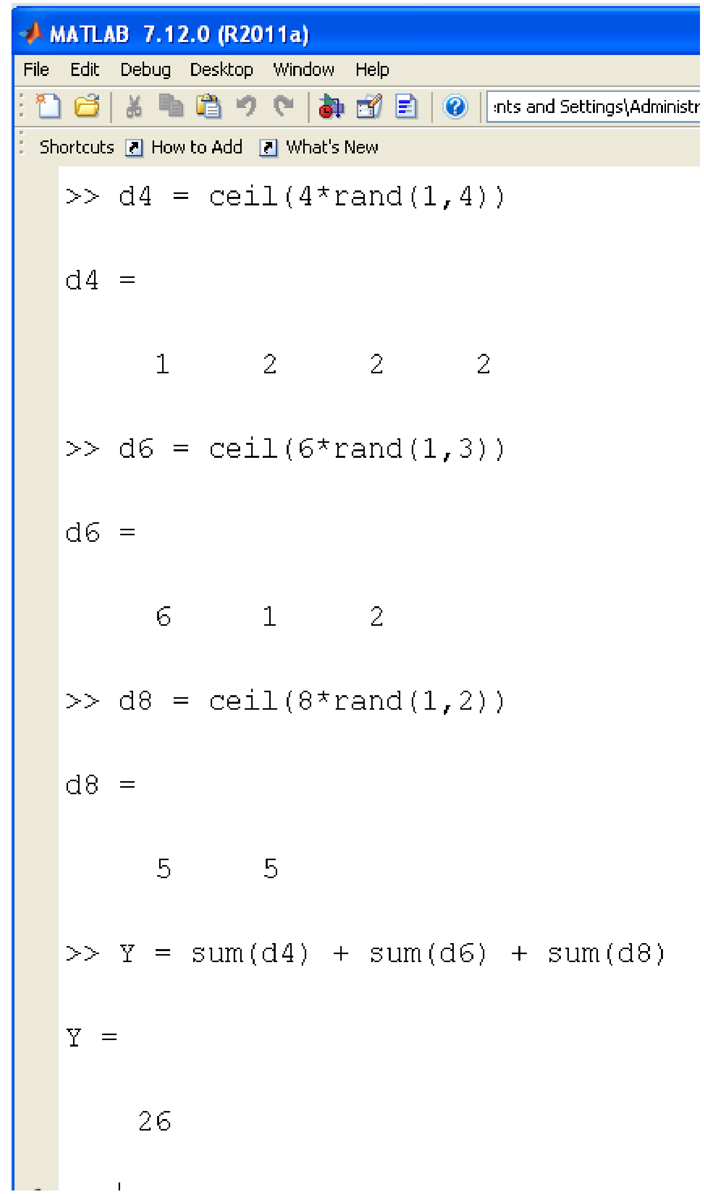
Example: Dice

Let

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

Determine the probability:

- p(y = 35)
- p(y > 35)
- The 90% confidence interval (5% tails)

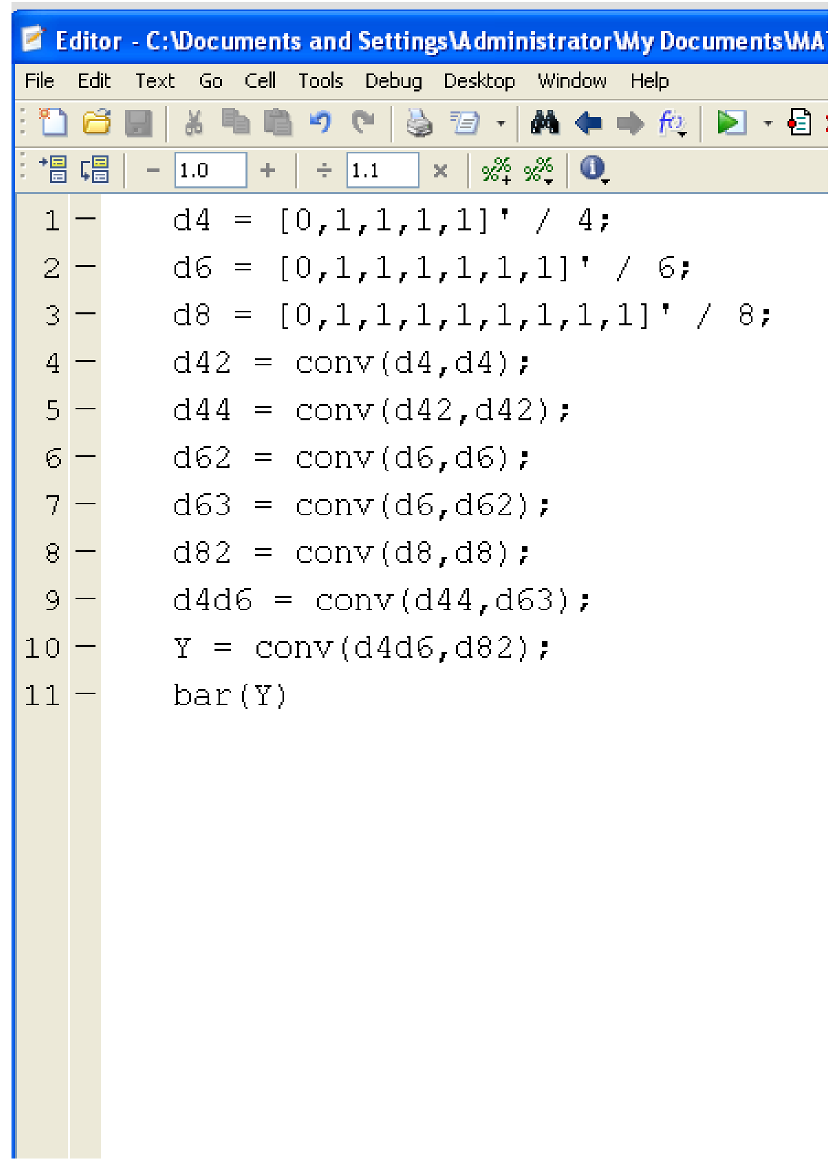


Solution #1: Enumeration

- Y = 4d4 + 3d6 + 2d8

Test every possible outcome

- Nested for-loops works
- Convolution works

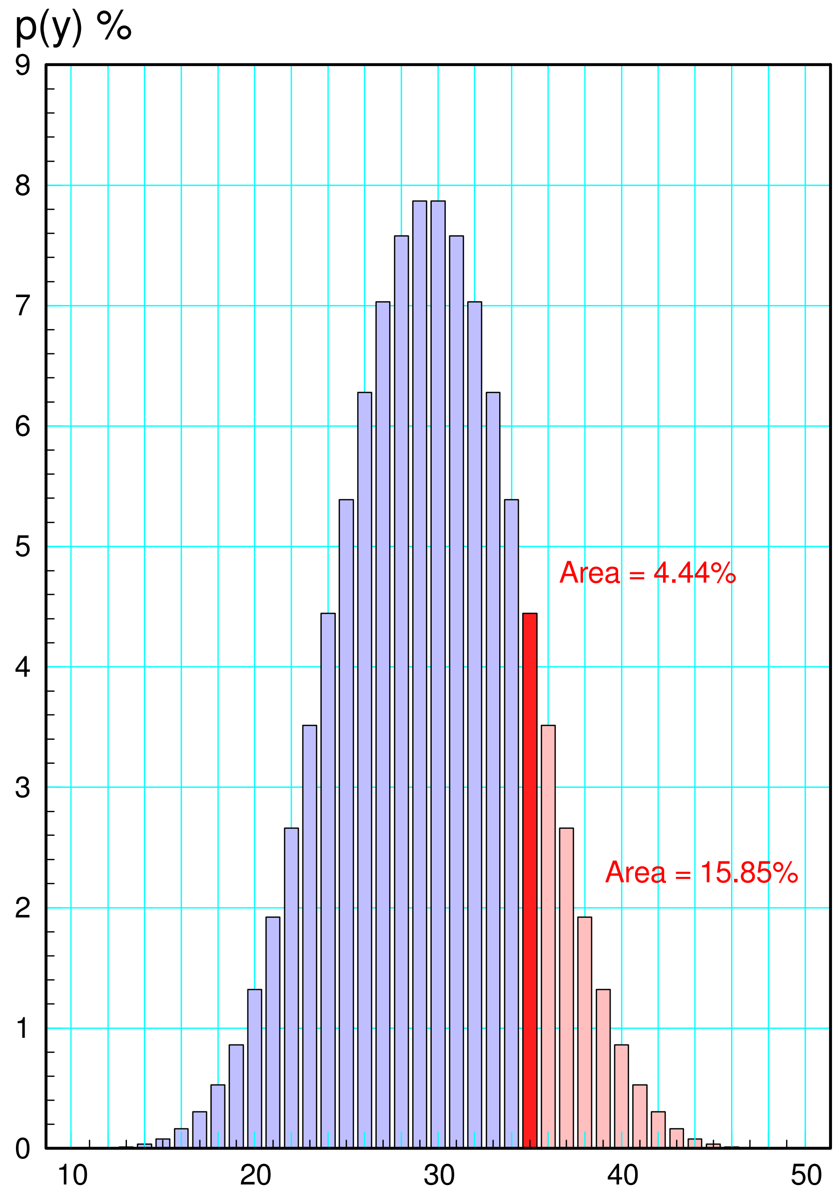


Enumeration Results

- Results are exact (good)
- p(y=35) = 4.4445%
- p(y ≥ 35) = 15.8524%
- 90% confidence interval: 21.5 < y <38.5

Problem with enumeration:

- All 3,538,944 outcomes tested
- If each test costs $10, that's $35 million



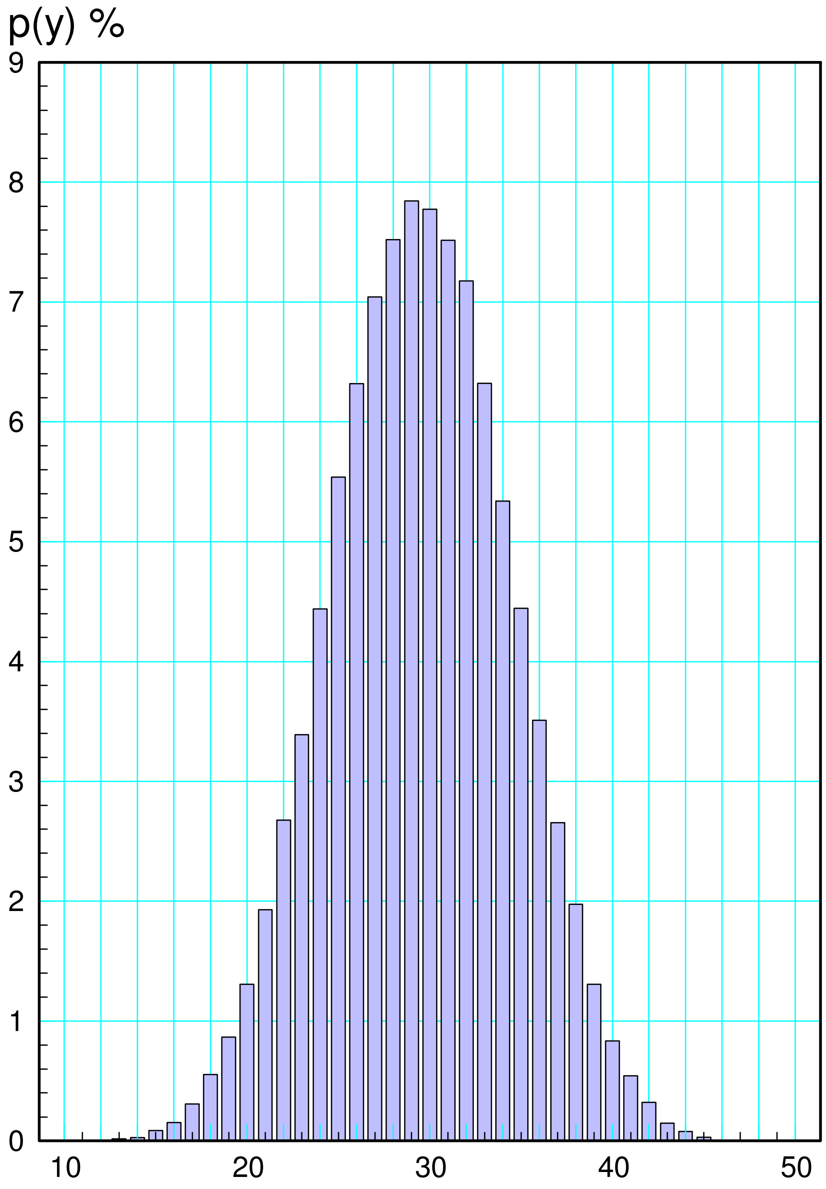
Solution #2: Monte-Carlo

Roll the dice 100,000 times

Record the frequency of each result

Matlab Code

```
RESULT = zeros(52,1);
for n=1:1e5
   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);
   RESULT(Y) = RESULT(Y) + 1;
   end
bar(RESULT)
```



Monte-Carlo Results

p(y = 35) = 4.444%

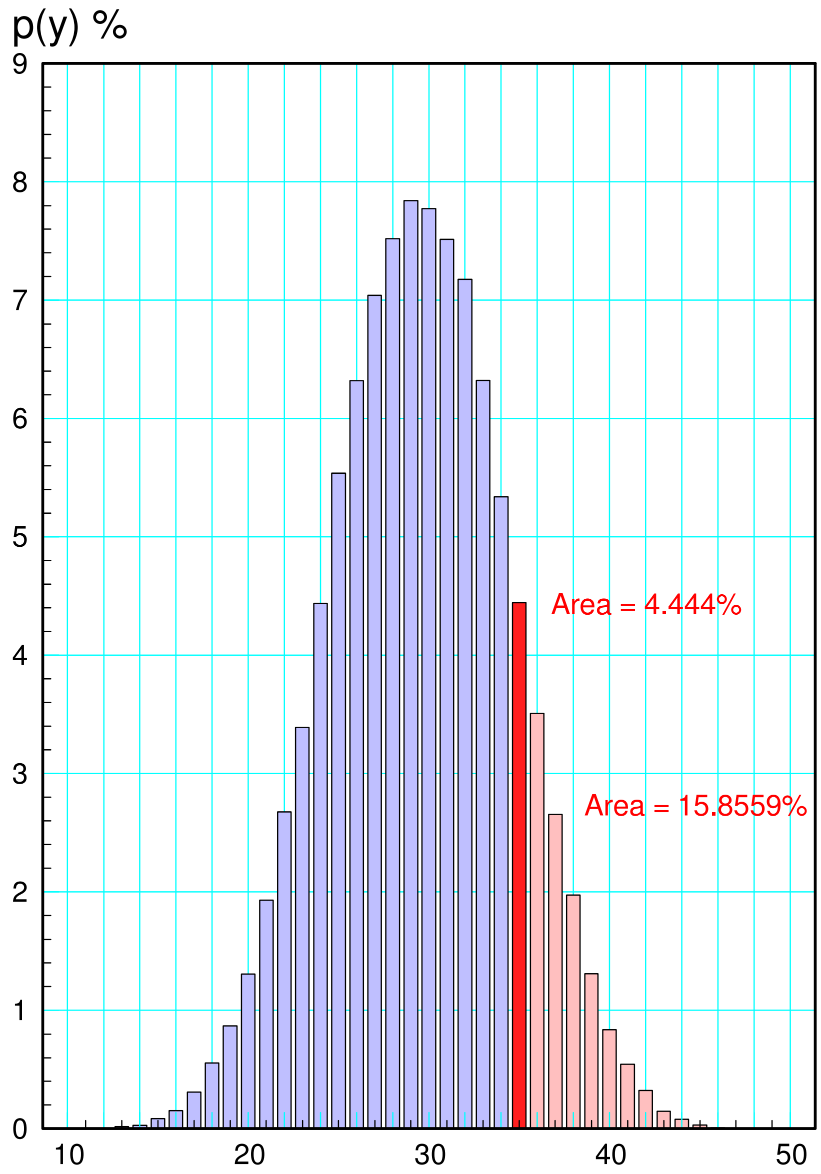
- exact = 4.4445%

p(y >= 35) = 15.859%

- exact = 15.8524%

90% confidence interval: 20.5 < y < 38.5

- 21.5 < y < 38.5



Problem with Monte-Carlo Simulations

Expense

- If it costs $10 for each experiment, 100,000 samples = $1 million

Time

- If it takes 10 minutes to measure each y, 100,000 samples = 1.9 years

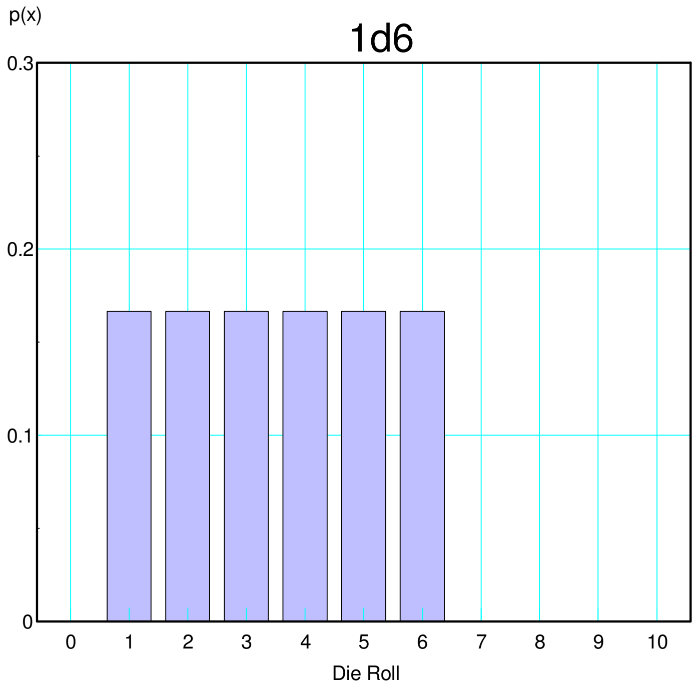
Can you come up with the same results using fewer measurements?

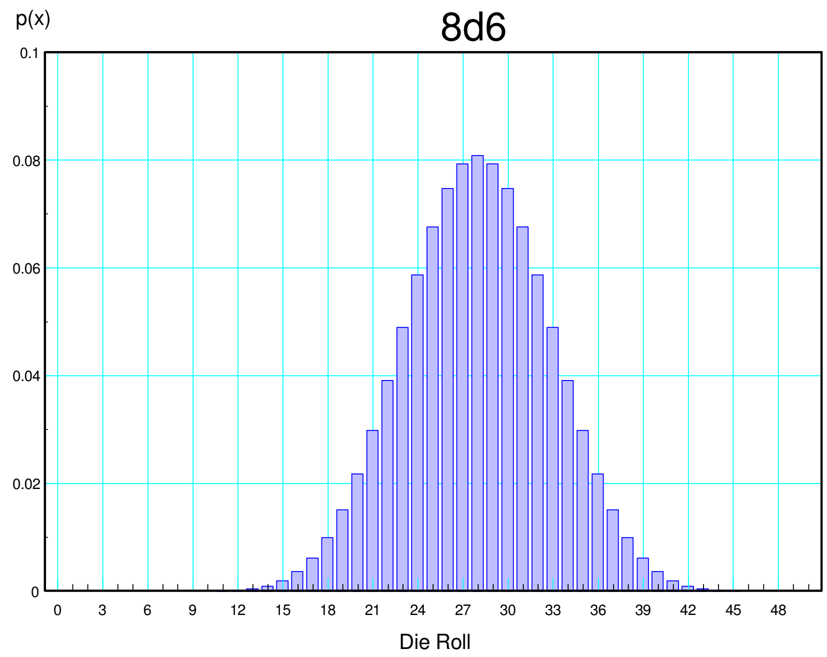
- Yes
- Requires statistics

Central Limit Theorem

- All distributions converge to a normal distribution
- Normal + Normal = Normal
- Once you have a normal distribution, you remain with a normal distribution.

For example, 1d6 (not normal) vs. 8d6 (approx normal)

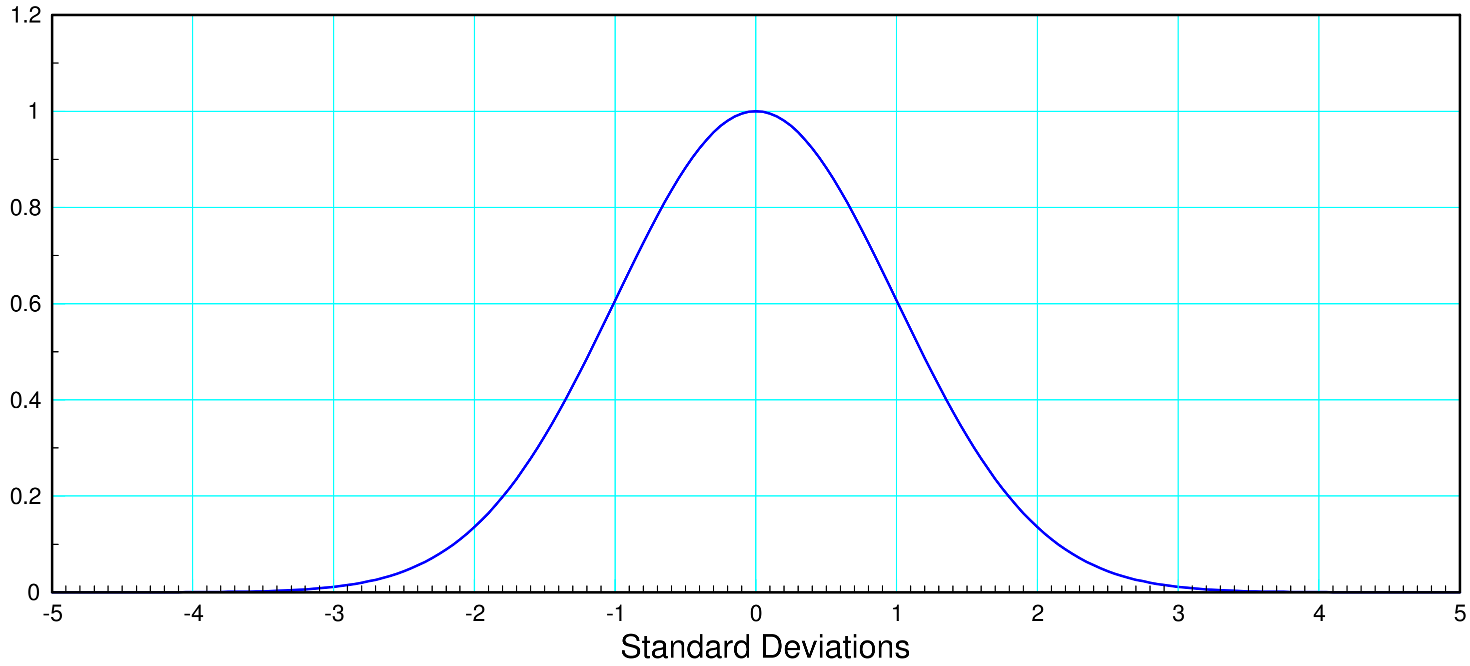




Normal (Gaussian) Distributions:

$$N(\bar{x}, \sigma^2)$$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(\frac{-(x-\bar{x})^2}{2\sigma^2}\right)$$



Standard Normal Distribution (normalized so the peak is 1.000)

Properties of Normal Distributions

Two parameters define a normal distribution

- Mean
- Standard Deviation (or Variance)

Mean: average of the data

$$\mu = \frac{1}{n} \sum y_i$$

Variance: average squared distance to mean

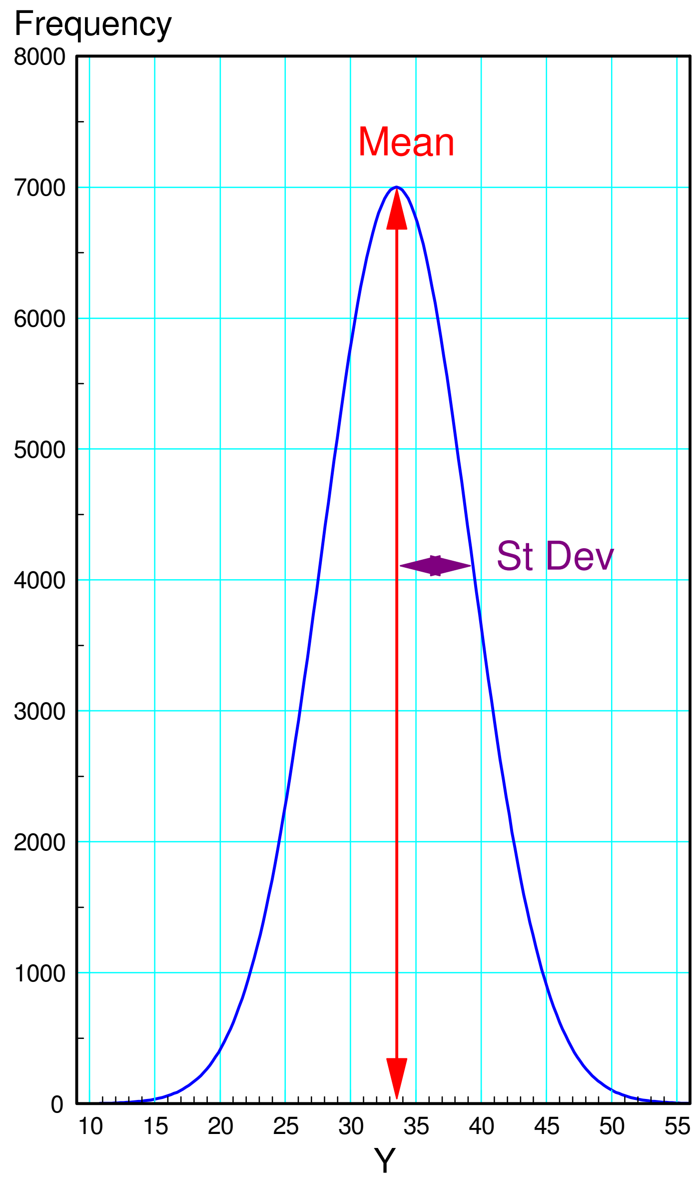
$$\sigma^2 = \frac{1}{n} \sum (y_i - \mu)^2$$

Standard Deviation (spread)

$$\sigma = \sqrt{\sigma^2}$$

When you add normal distributions

- The means add
- The variances add



Example: Dice

The mean and variance for a 4-sided die

```
>> d4 = [1,2,3,4];
>> m4 = sum(d4) / 4
        m4 =    2.5000
>> v4 = sum( (d4 - m4).^2 )/4
        v4 =    1.2500
```

The mean and variance for a 6-sided die

```
>> d6 = [1,2,3,4,5,6];
>> m6 = sum(d6) / 6
        m6 =    3.5000
>> v6 = sum( (d6 - m6).^2 )/6
        v6 =    2.9167
```

The mean and variance for an 8-sided die

```
>> m8 = sum(d8) / 8
m8 =    4.5000
>> v8 = sum( (d8 - m8).^2 )/8
v8 =    5.2500
```

4d4 + 3d6 + 2d8

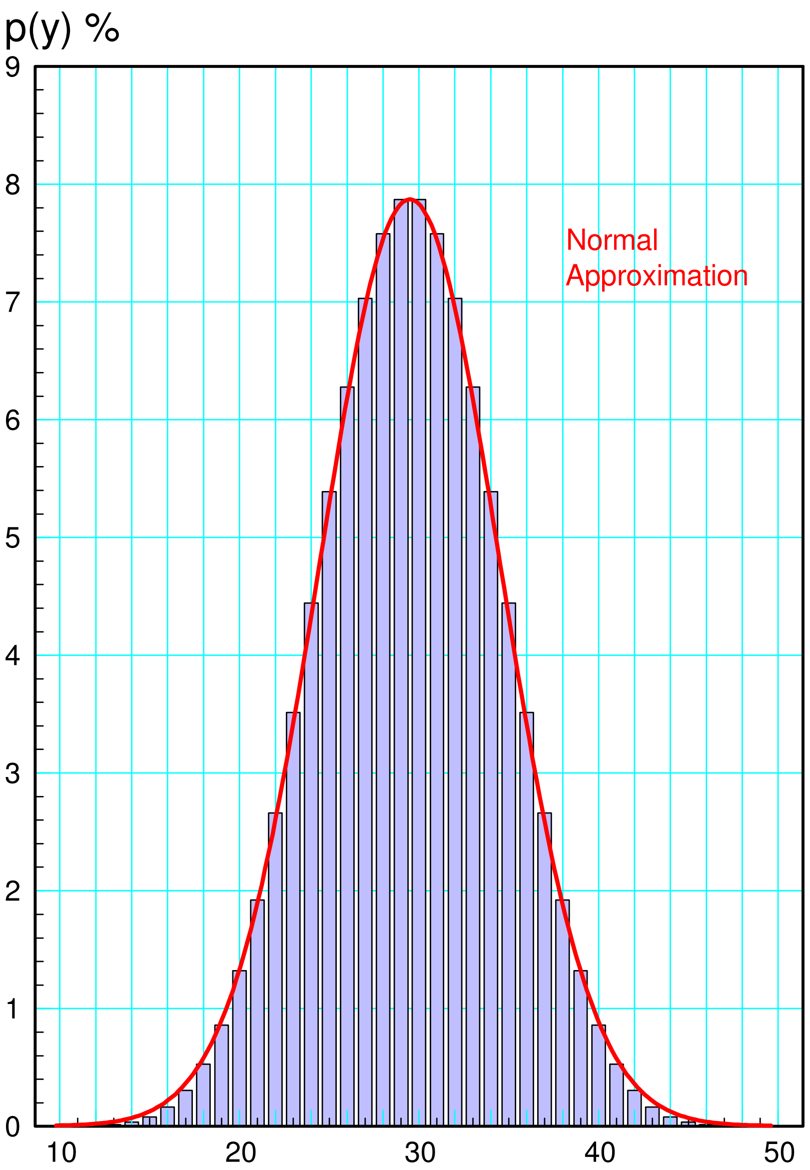
- The means add
- The variances add

```
>> my = 4*m4 + 3*m6 + 2*m8
my =   29.5000

>> vy = 4*v4 + 3*v6 + 2*v8
vy =   24.2501

>> sy = sqrt(vy)
sy =    4.9244
```

To plot the normal pdf

```
s = [-4:0.01:4]';
p = exp(-s.^2 / 2);
plot(s*sy+my,p*7000);
xlabel('Die Roll')
```

Central Limit Theorem in action...



What is the probability of y = 35?

- $p(34.5 < y < 35.5)$

Calculate the z-score

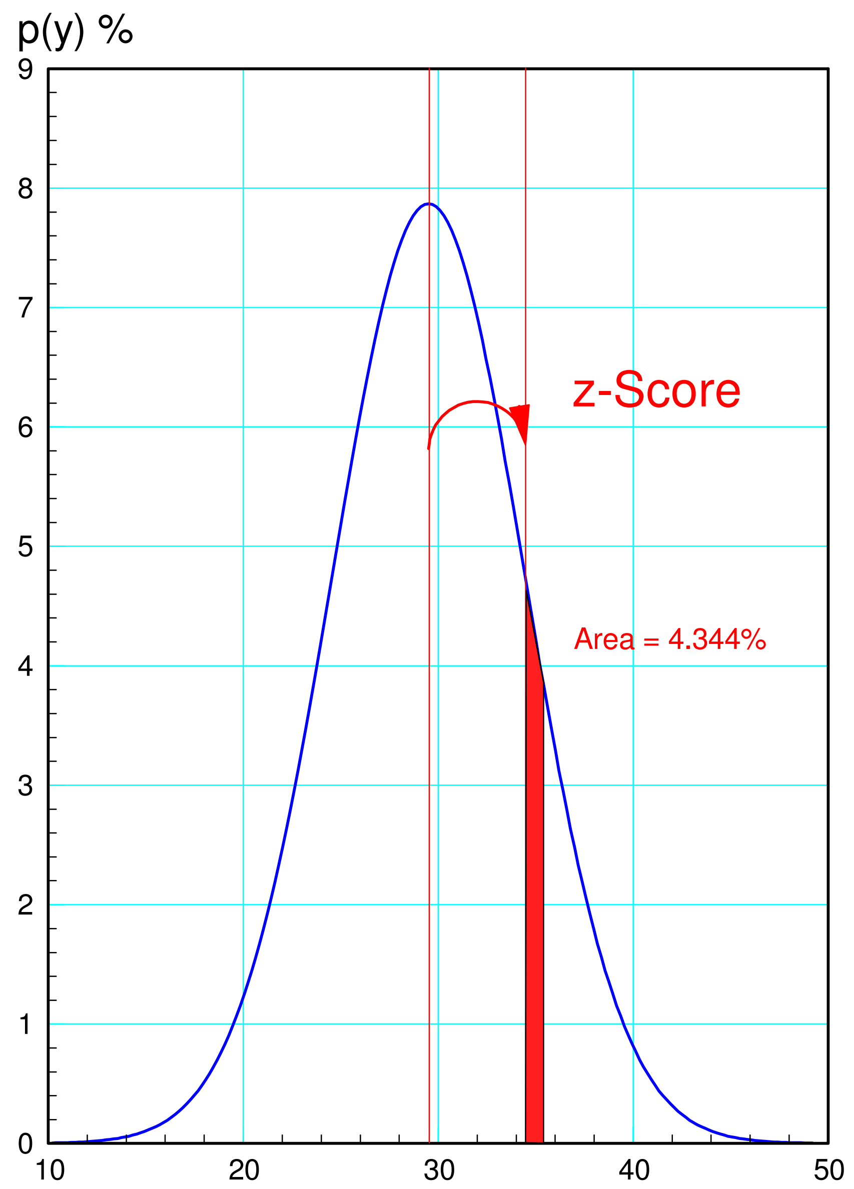
- how far y is from the mean

$$z_1 = \left(\frac{34.5-\mu}{\sigma}\right)$$

```
>> z1 = (34.5 - my)/sy
z1 =    1.0153
p(y > 34.5) = 15.498% (from StatTrek)

>> z2 = (35.5 - my)/sy
z2 =    1.2184
p(y > 35.5) = 11.154% (from StatTrek)
```

Take the difference

- $p(34.5 < y < 35.5) = 4.344\%$
- exact = 4.4445%

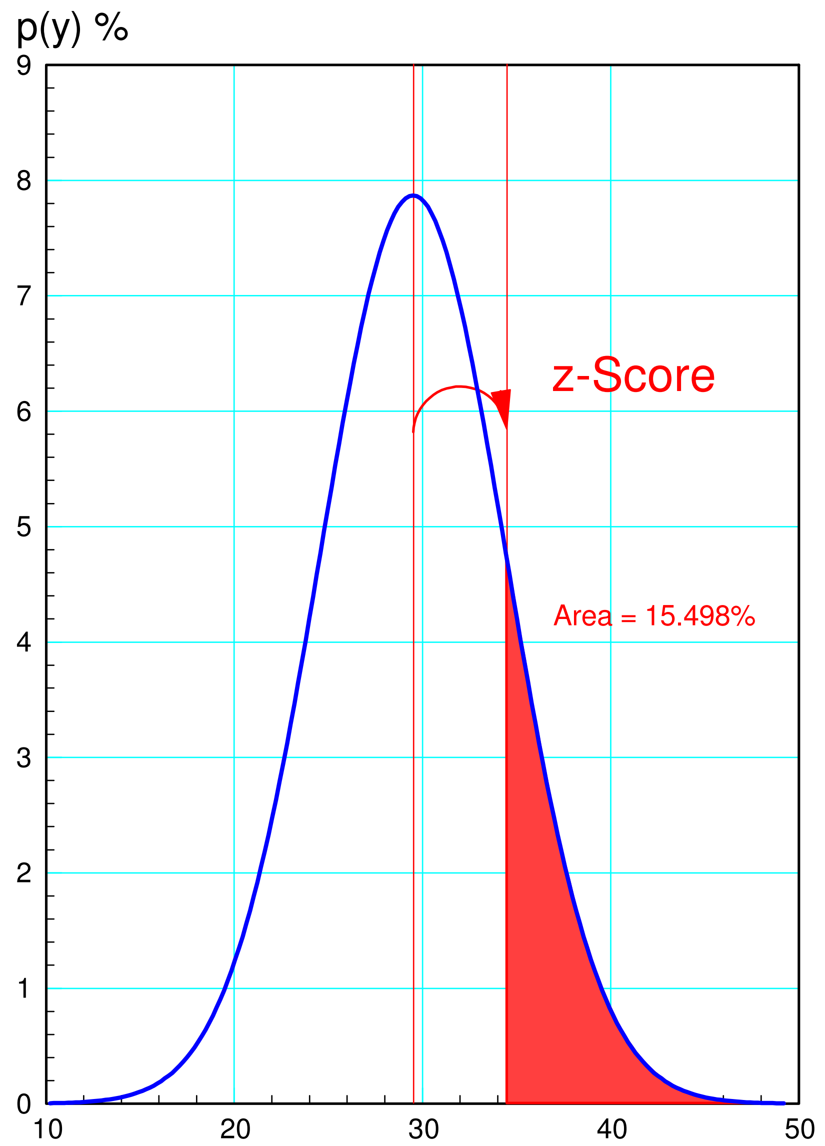


What is the probability that y > 35?

- Find the area to the right of 34.5

```
>> z1 = (34.5 - my)/sy
z1 =     1.0153
```

- p = 15.498% (from StatTrek)
- p = 15.8524% (exact)



Standard Normal Table

- z-score = 1.0565
- p = 14.573%

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47209	0.46812	0.46415
0.1	0.46017	0.45620	0.45224	0.44829	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41684	0.41294	0.40904	0.40516	0.40130	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32635	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29805	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25784	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23269	0.22965	0.22663	0.22363	0.22065	0.21769	0.21476
0.8	0.21186	0.20897	0.20611	0.20327	0.20046	0.19766	0.19489	0.19215	0.18943	0.18673
0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17105	0.16853	0.16602	0.16354	0.16109
1.0	0.15865	0.15625	0.15387	0.15150	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10384	0.10204	0.10027	0.09852
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07214	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592

What is the 90% confidence interval?

- Two tails, each 5%

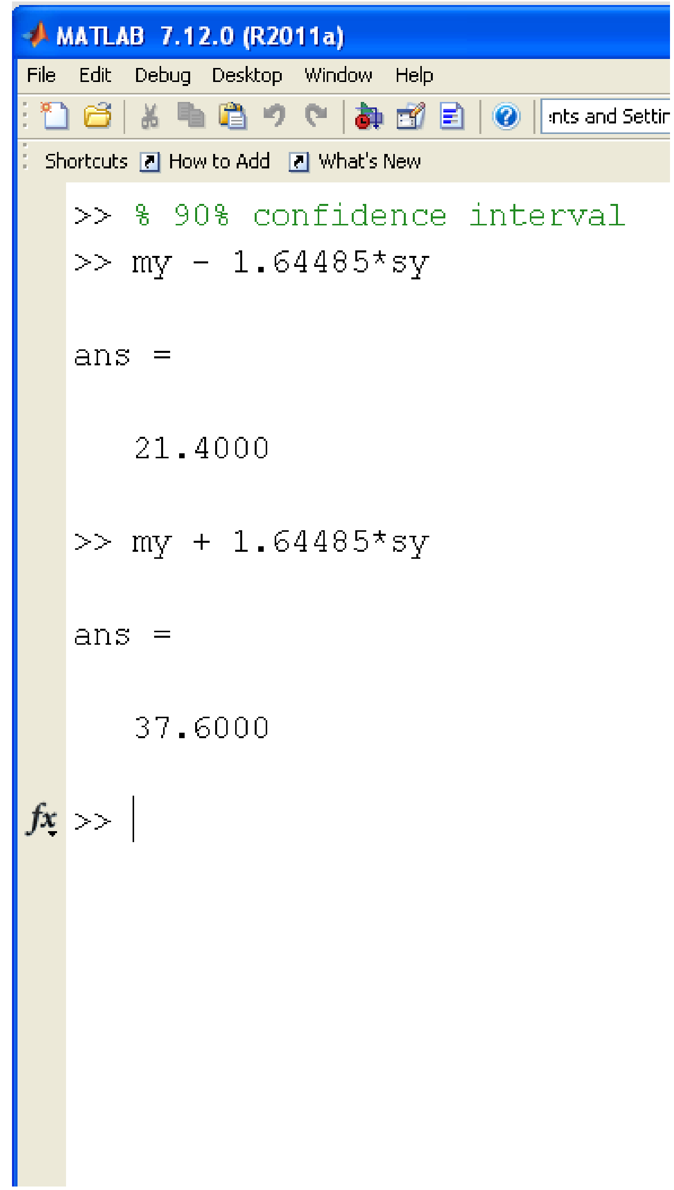
What is the z-score for 5% tails?

- z-score = 1.64485
- StatTrek works
- Standard Normal table works

	0.03	0.04	0.05	0.06
1.5	0.06301	0.06178	0.06057	0.05938
1.6	0.05155	0.05050	0.04947	0.04846
1.7	0.04181	0.04093	0.04006	0.03920

Result

- $21.4 < y < 37.6$ (p = 90%)
- $21.5 < y < 38.5$ (from enumeration)



Comment

If you know the mean and the standard deviation, you can calculate the odds

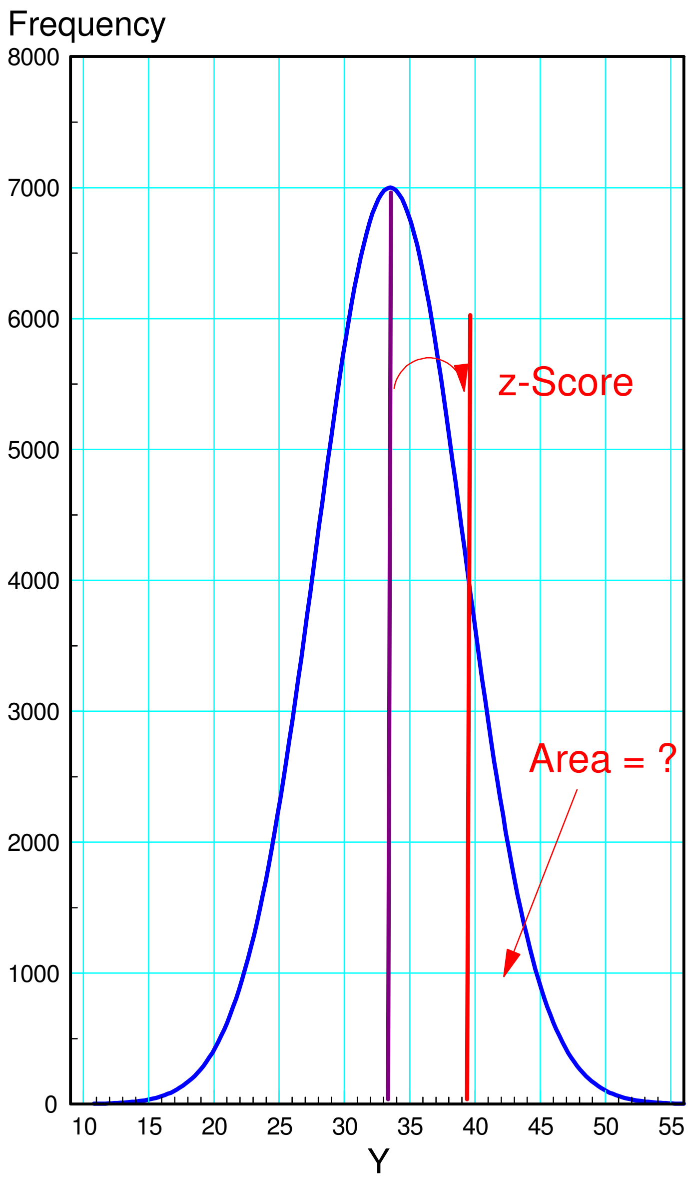
- Find the z-score
- Distance to the mean in terms of standard deviations

Convert z-scores to probabilities

- Standard-Normal Table
- StatTrek

This requires zero die rolls

- Saving a lot of money
- Saving a lot of time



Problem: What if you *don't* know the mean and standard deviation?

Solution:

- Collect some data
- Estimate the mean and standard devitation from the data

The result is a Student-t Distribution

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

df \ p	0.001	0.0025	0.005	0.01	0.025	0.05	0.1	0.15	0.2
1	636.61900	318.30900	63.65670	31.82050	12.70620	6.31380	3.07770	1.96260	1.37640
2	31.59910	22.32710	9.92480	6.96460	4.30270	2.92000	1.88560	1.38620	1.06070
3	12.92400	10.21450	5.84090	4.54070	3.18240	2.35340	1.63770	1.24980	0.97850
4	8.61030	7.17320	4.60410	3.74690	2.77640	2.13180	1.53320	1.18960	0.94100
5	6.86880	5.89340	4.03210	3.36490	2.57060	2.01500	1.47590	1.15580	0.91950
6	5.95880	5.20760	3.70740	3.14270	2.44690	1.94320	1.43980	1.13420	0.90570
7	5.40790	4.78530	3.49950	2.99800	2.36460	1.89460	1.41490	1.11920	0.89600
8	5.04130	4.50080	3.35540	2.89650	2.30600	1.85950	1.39680	1.10810	0.88890
9	4.78090	4.29680	3.24980	2.82140	2.26220	1.83310	1.38300	1.09970	0.88340
10	4.58690	4.14370	3.16930	2.76380	2.22810	1.81250	1.37220	1.09310	0.87910
100	3.39050	3.17370	2.62590	2.36420	1.98400	1.66020	1.29010	1.04180	0.84520

Example: y = 4d4 + 3d6 + 2d8

Step 1: Collect Data

Roll the dice three times

```
DATA = [];
for i=1:3
   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);
   DATA = [DATA, Y];
   end
```

```
DATA =    30    40    28
```

Step 2: Compute the

- Mean
- Standard Deviation
- Sample Size

```
>> x = mean(DATA)
x =   32.6667

>> s = std(DATA)
s =    6.4291

>> n = length(DATA)
n =     3
```

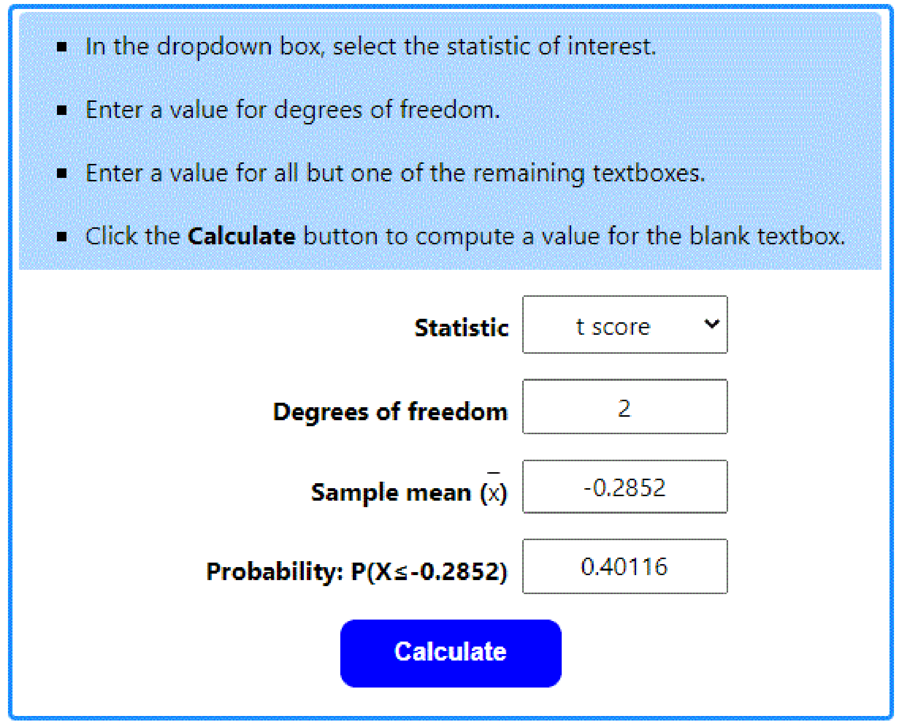
Probability 34.5 < y < 35.5

Calculate the distance to the mean

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

>> 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)
- p1 = 40.116%
- p2 = 35.124%



Difference = 4.992%

- vs. 4.4445% (exact)

Probability y > 34.5

Calculate the distance to the mean

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

Convert this to a probability using a Student-t table

- StatTrek also works
- Degrees of Freedom = 2 (Sample Size - 1)
- p1 = 40.116%
- exact = 15.8524%

90% Confidence Interval

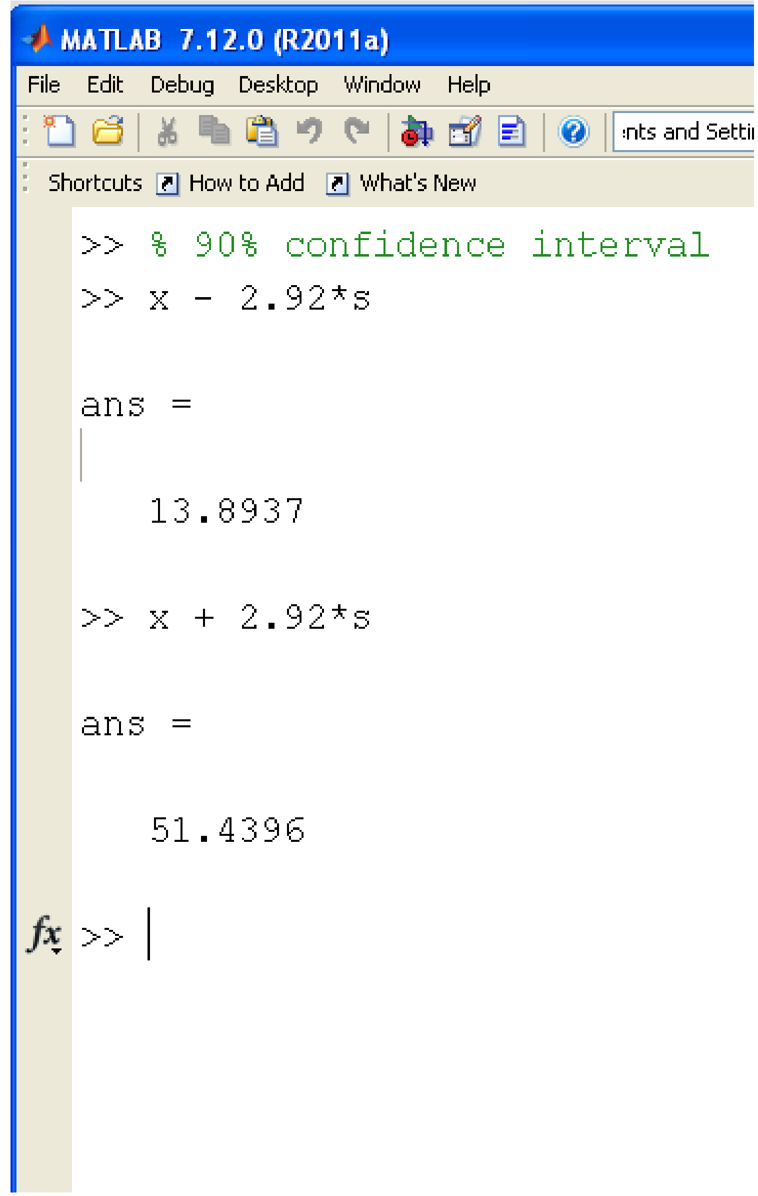
Determine the t-score for

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

Go left and right of the mean by 2.92 standard deviations

Result

- 13.89 < y < 41.54 (p = 90%)
- 21.5 < y < 38.5 (from enumeration)



Example: Roll Ten Dice

- y = 4d4 + 3d6 + 2d8

Step 1: Collect Data

```
DATA = [];
for i=1:10
   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);
   DATA = [DATA, Y];
   end
```

```
DATA =

    30    21    32    36    29    24    27    35    20    35
```

Step 2: Compute the

- Mean
- Standard Deviation
- Sample Size

```
>> x = mean(DATA)
```

x = 28.9000

```
>> s = std(DATA)
```

s = 5.8205

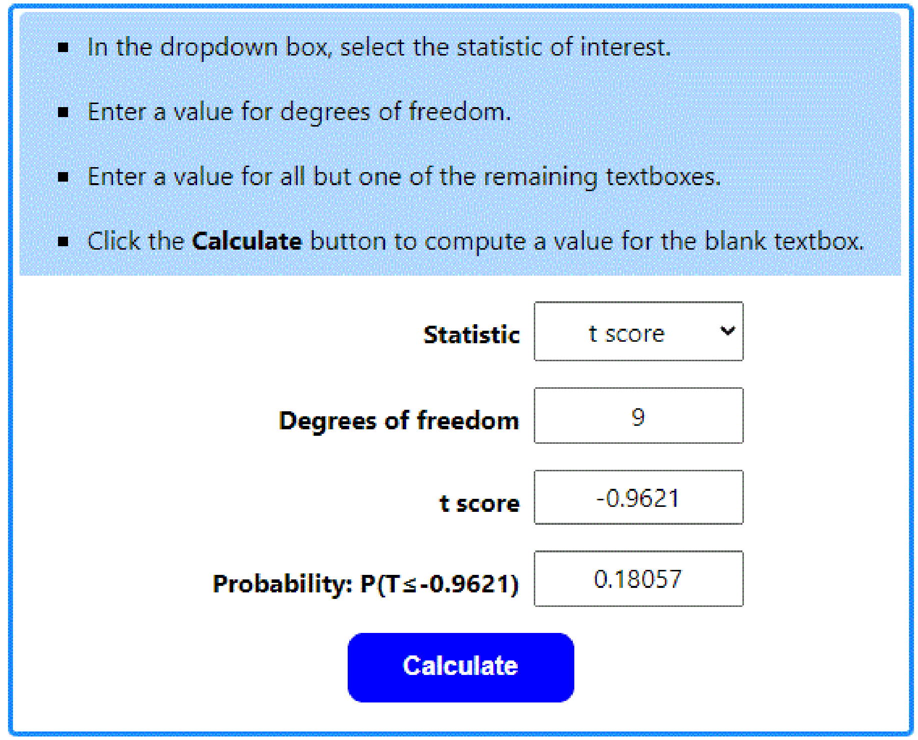
```
>> n = length(DATA)
```

n = 10

Probability 34.5 < y < 35.5

Calculate the distance to the mean

```
>> t1 = (34.5 - x) / s
t1 =     0.9621

>> t2 = (35.5 - x) / s
t2 =     1.1339
```



Convert this to a probability using a Student-t table

- StatTrek also works
- Degrees of Freedom = 9 (Sample Size - 1)
- p1 = 18.057%
- p2 = 14.307%

Difference = 3.75%

- vs. 4.4445% (exact)

Probability y > 34.5

Calculate the distance to the mean

```
>> t1 = (34.5 - x) / s
t1 =    0.9621
```

Convert this to a probability using a Student-t table

- StatTrek also works
- Degrees of Freedom = 9 (Sample Size - 1)
- p1 = 18.057%
- exact = 15.8524%

90% Confidence Interval

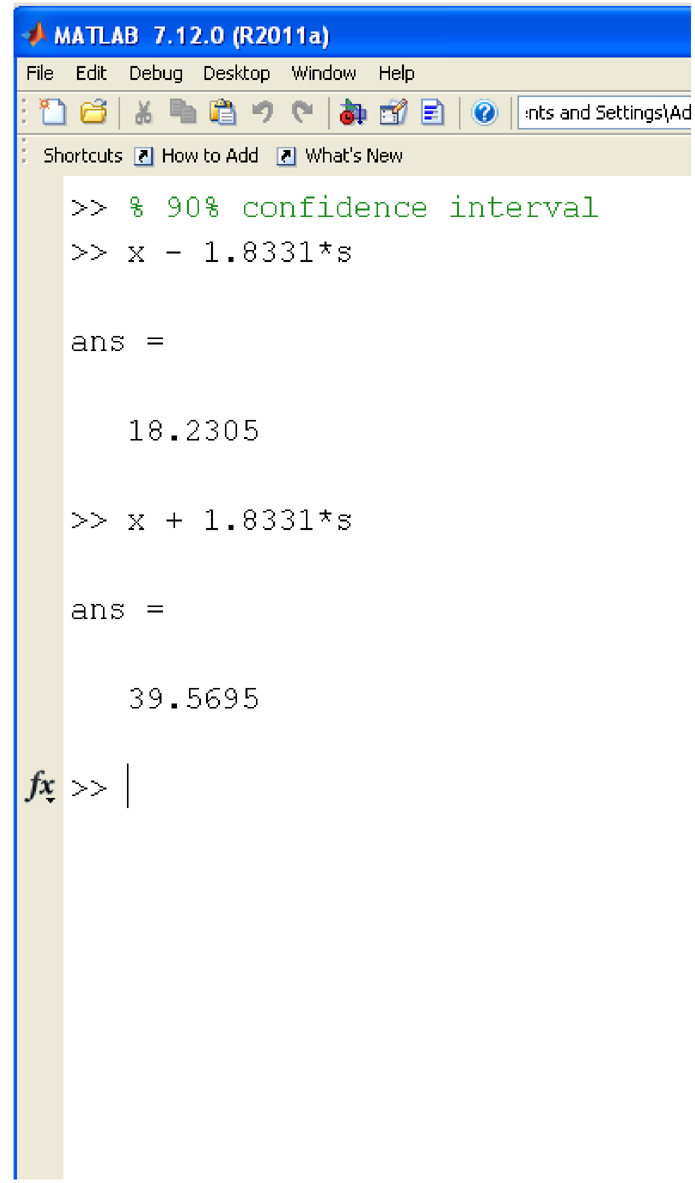
Determine the t-score for

- 9 degrees of freedom
- 5% tails
- $t = 1.8331$

Go left and right of the mean by 1.8331 standard deviations

Result

- $18.23 < y < 39.56$ ($p = 90\%$)
- $21.5 < y < 38.5$ (from enumeration)



Repeat Using 30 Rolls

```
>> x = mean(DATA)
x =   29.6333

>> s = std(DATA)
s =    4.4527

>> n = length(DATA)
n =    30

>> t1 = (34.5 - x) / s
t1 =    1.0930
p1 = 14.17% (from StatTrek)

>> t2 = (35.5 - x) / s
t2 =    1.3176
p2 = 9.898% (from StatTrek)
```

Results: Y = 4d4 + 3d6 + 2d8

	# 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 sample size = 3	3	4.992%	40.116%	(13.9, 41.5)
t-Test sample size = 10	10	3.75%	18.057%	(18.2, 39.5)
t-Test sample size = 30	30	4.27%	14.17%	(22.1, 37.2)

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