Student t Distribution

Probability and hypothesis testing

Introduction

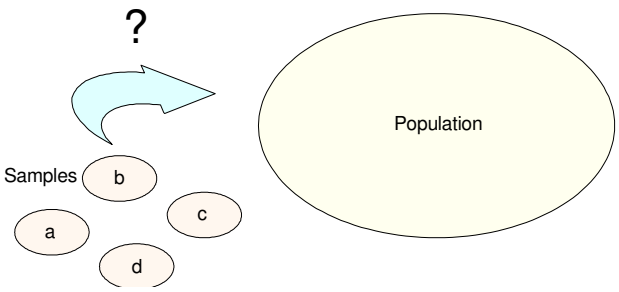
Assume you have a normally distributed random variable with a known mean and variance. Due to the *central limit theorem*, the assumption of a normal distribtion is fairly broad: many if not most of the random variables you encounter will have a normal distribution. The assumption that you *know* the mean and variance is also not terribly constraining. Many times we *do* know the mean and variance. For example,

- When summing dice, you know precisely what the mean and variance for a single die are. Likewise, you can compute the mean and variance of the sum of various combinations of dice.
- When summing uniform distibutions, you know precisely what the mean and variance are. You likewise know the mean and variance for the sum of various uniform distributions.

If you know the mean and variance, you can analyze various probabilities using a normal distribution.

Here's the problem though. Suppose you have a population which is normally distributed but you do *not* know the mean and variance. Suppose you collect N samples from this population. How to you analyze this data?

This is where the Student t Distribution comes into play. With it, you can analyze a finite number of samples from a normally distibuted population.



Problem for a t-Distribution: Determine probabilities from a population using a finite sample size

A student t-test is a test of a mean. The heart of a t-test is the Central Limit Theorem. This states that all distributions converge to a Normal distribution (the bell-shaped curve you're probably familiar with). Furthermore, if you add a Normal distribution to a Normal distribution, the result is a Normal distribution.

In this lecture, we'll

- Review methods to find probabilities related to rolling dice using enumeration and Monteo Carlo experiments
- Review methods to find these probabilities using a Normal approximation, and then
- Cover how to find these probabilities using only a finite number of results from rolling the dice.

The latter uses a *Student t-Distribution* and is called *a t-Test*.

Probabilities with Dice: Monte Carlo

Just to give us something concrete to focus only, assume Y is the sum of four 4-sided dice , three 6-sided dice, adnd 2 8-side dice

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

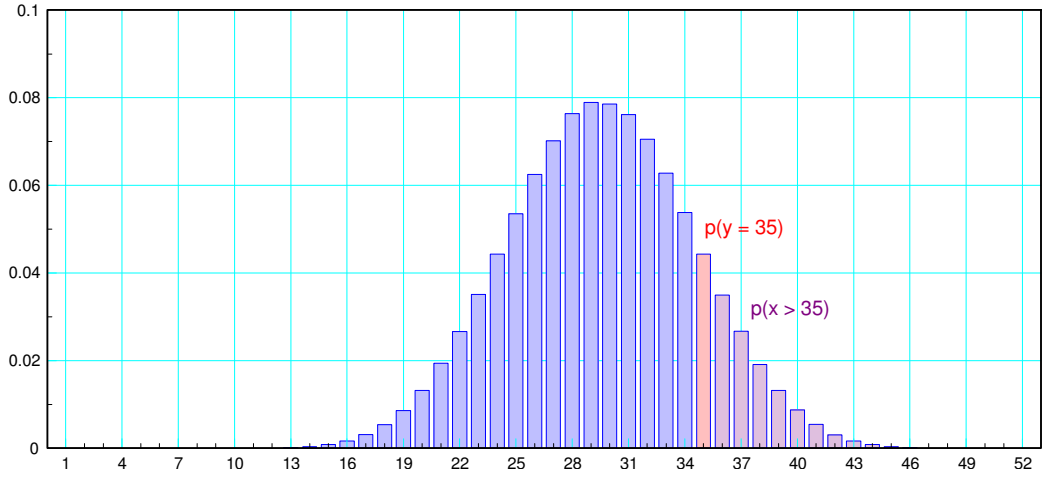
Determine:

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

The first approach we covered in this course used Monte Carlo experiments. With this method, you roll the dice a large number of times and count the number of occurences:

```
RESULT = zeros(52,1);
for n=1:1e6
   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
pdf = RESULT / 1e6'
bar(pdf)
```

The result for 1 million rolls are as follows:



pdf for Y = 4d4 + 3d6 + 2d8

The probabities we're looking for are thus (from Matlab:

The probabilty that y = 35 is 4.43%

```
pdf(35)
ans =    0.0443
```

The probability that y > 35 is 11.39%

```
>> sum(pdf(36:52))
ans =    0.1139
```

The right limit for 5% tails is approximately 38:

```
>> sum(pdf(38:52))
ans =    0.0523
```

The left limit for 5% tails is approximately 21

```
>> sum(pdf(1:21))
ans =    0.0526
```

The 90% confidnce interval is (22, 37)

Probabilities with Dice: Normal Approximation

One problam with Monte-Carlo experiment is this took one million die rolls. If it costs $1 each time you roll the dice, you just blew $1 million running this experiment.

A second way to find these probabilities is to use a Normal approximation. With dice, we *know* the mean and variance for each die. For a uniform distribution

$$\mu = \left(\frac{b+a}{2}\right)$$

$$\sigma^2 = \left(\frac{(b-1+a)^2-1}{12}\right)$$

When summing normal distributions

- The means add and
- The variance adds

This lets us compute the mean and variance of 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

Once you know the mean and variance, you can compute odds using a Normal table and z-scores.

p(y = 35): Since this is a continuous distribution, you cant find the area of a point (it will be zero). Instead, find the area in the interval (34.5, 35.5). In Matlab

```
>> 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
```

The probability that y > 35

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

The 90% confidence interval

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

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

Note that the results match up very closely to the Monte Carlo results. These calculations required zero rolls, however. If each roll costs $1, this saves $1 million dollars.

	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

Student t-Ditribution

In order to use the normal approximation, you need to know the populations mean and variance. Suppose you *don't* know what these are. Then what do you do?

If you don't know the population's mean and variance, you can estimate these using n samples from the poplation. The mean and variance are then compute *almost* the same way you would for a Normal distribution.

If you measure *every* member of a population, the mean and variance are

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

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

The z-score is then the distance from a point to the mean in terms of standard deviations

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

If, on the other hand, you estimate the mean and variance from a sample from the population, the mean and variance are computed slightly differently:

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

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

The distance from the mean is now called *the t-score*

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

In addition the number of data points is important. This determines another paramater: *the degrees of freedom*

$$dof = n - 1$$

Note that when computing the t-score, you are dividing

- A normal distribution (computed value of the mean), by
- A gamma distribution (computed value of the standard deviation).

This ratio has a Student t-Distribution (or t-Distribution for short).

A t-Distribution looks very much like a Normal Distribution - except that it takes sample size into account.

df \ p	0.001	0.0025	0.005	0.01	0.025	0.05	0.1
1	636.619	318.309	63.6567	31.8205	12.7062	6.3138	3.0777
2	31.5991	22.3271	9.9248	6.9646	4.3027	2.92	1.8856
3	12.924	10.2145	5.8409	4.5407	3.1824	2.3534	1.6377
4	8.6103	7.1732	4.6041	3.7469	2.7764	2.1318	1.5332
5	6.8688	5.8934	4.0321	3.3649	2.5706	2.015	1.4759
6	5.9588	5.2076	3.7074	3.1427	2.4469	1.9432	1.4398
7	5.4079	4.7853	3.4995	2.998	2.3646	1.8946	1.4149
8	5.0413	4.5008	3.3554	2.8965	2.306	1.8595	1.3968
9	4.7809	4.2968	3.2498	2.8214	2.2622	1.8331	1.383
10	4.5869	4.1437	3.1693	2.7638	2.2281	1.8125	1.3722
100	3.3905	3.1737	2.6259	2.3642	1.984	1.6602	1.2901

Student t Table

With a Student t-Table,

- The left column is the degrees of freedom. This is the sample-size minus one.
- The top is the area of the tail

The number is the middle is the corresponding t-score - similar to the z-score for a normal distribution. As the sample size goes to infinity, the t-table converges to a Normal distribution. A t-table is also available on StatTrek.

Essentially, a t-Table is a z-table, only the sample size is taken into account. For example, if you want 5% tails,

- The z-score is 1.64485,
- The t-score with infinite degrees of freedom is also 1.64485,
- The t-score with 10 degrees of freedom increases to 1.8125,
- The t-score with 5 degrees of freedom increses to 2.5706, and
- The t-score with 1 degree of freedom increases to 6.3138

What's happening is you're just being more and more cautious as the sample size gets smaller. To make up for the lack of information, the confidence intervals widen for a given level of probability.

t-Score with a Sample Size of One

Going back to our example of

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

Determine the

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

assuming you do *not* know the mean and variance of Y. Instead, you collect a single measurement:

```
DATA = [];
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);
   DATA = [DATA, Y];
   end

DATA = 30
```

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-Score with a Sample Size of 3

Assume you collect three data points:

```
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
```

Now, you *can* actually get results. The mean, variance, and sample size can be computed using Matlab (note the Matlab commands var() and std() actually use n-1 in their calculations).

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

>> s = std(DATA)
s =    6.4291

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

Using a t-Table, you can now compute various probabilities.

Probability 34.5 < y < 35.5: Calculate the distance to the mean

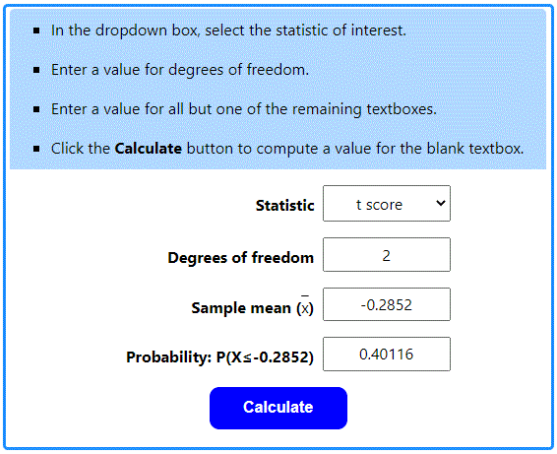
```
>> t1 = (34.5 - x) / s
t1 =    0.2852

>> 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)



Coverting a t-Score to a probability using StatTrek

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)

Note that these resuls are in the right ballpark - but are off a bit. That's not too surprising since the sample size is only three.

	Monte-Carlo	Normal Approx	t-Test
p(y = 35)	4.43%	4.34%	4.992%
p(y > 35)	11.39%	11.15%	40.11%
90% confidnce interval	22 < y < 37	21.4 < y < 37.6	13.89 < y < 41.54
# rolls	1,000,000	0	3

Results for a t-Test with a sample size of 3

t-Score with a Sample Size of 10 and 30

If you increase the sample size, the results get better:

- With more data, the mean and variance approach the true mean and variance of the population, and
- The t-scores get smaller as sample size goes up

	# 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%	40.116%	(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)

t-Tests for Populations vs. Individuals

The way you calculate the variance for a t-test depents upon what question you're asking.

If you're asking a question about an individual, such as *what's the value of a single die roll?,* the variance is computed as

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

This is the function *var()* in Matlab.

```
s2 = var(Data);
```

If you're asking a question about a population, such as *what the average of Y?* the variance is reduced by the sample size:

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

or in Matlab

```
s2 = var(Data) / n;
```

What's happening is you know more about populations than you know about individuals.

- For a single roll of the dice, the variance approaches a constant as the sample size goes to infinity.
- For populations, the variance goes to zero as the sample size goes to infinity.

As I collect more and more data, I know more and more about the population's mean. In the limit, I know precisely what the population's mean is.

Whether you do or do not divide by n depends upon the question you're asking:

- Is it about an individual measurement? (you don't divide by n), or
- Is it about a population's average? (you do divide by n).

For example, suppose I want to know the 90% confidence interval for the average of Y. Since I want to know the population's average, you divide the variance by the sample size.

```
x = mean(Data);
v = var(Data) / n;
s = std(Data) / sqrt(n);
```

the 90% confidnnc interval is then

$$x = \bar{x} \pm t \cdot s$$

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	29.5

Summary

The signifigance of a Student t-Test is you can determine probabilities related to a population using only a small number of measurements. This is important. Suppose, for example, you want to know something about the product your company is making.

- If you measure nothing, you have no idea what your selling.
- If you measure everyting, you go broke. Measuring everyting costs money. Worse, your entire product line is now used.

If you just measure a small sample of your product, however, you can estimate the mean and variance of your product. The resulting distribution is a Student t-Distribution. This looks very much like a Normal distribution, but it takes sample size into account.