# Chi-Squared Test

## ECE 341: Random Processes

### Lecture #25

note:  All lecture notes, homework sets, and solutions are posted on www.BisonAcademy.com

# Chi-Squared Test

Is the data consistent with an assumed distribution. It is used to test

- Whether a die is fair (each number has equal probability)
- Whether a distribution is Normal (vs. Poisson or geometric)

The Chi-Squared distribution is a type of Gamma distribution:

$$(x_i - \mu)^2$$

## Example: Is this a fair die?

```
d6 = ceil(rand*6);
```

## To run a Chi-Squared test to see if this really is a fair die, do the following:

- Divide the results into M bins (6 bins in this case: numbers 0 .. 5)
- Collect n data points.
- Count how many times the data fell into each of the M bins
- Compute the Chi-Squared total for each bin as

$$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$$

  - *np is the expected number of times data should fall into each bin*
  - *N is the actual number of times data fell into each bin*
- Use a Chi-Squared table to convert the resulting Chi-Squared score to a probability. Note that the degrees of freedom is equal to the number of bins minus one.

# Example: n = 120 die rolls

```
RESULT = zeros(1,6);

for i=1:120
    d6 = ceil(rand*6);
    RESULT(d6) = RESULT(d6) + 1;
end
```

| Die Roll (bin) | p theoretical probability | np expected frequency | N actual frequency | $\chi^2 = \left(\frac{(np-N)^2}{np}\right)$ |
|---|---|---|---|---|
| 1 | 1/6 | 20 | 18 | 0.2 |
| 2 | 1/6 | 20 | 27 | 2.45 |
| 3 | 1/6 | 20 | 25 | 1.25 |
| 4 | 1/6 | 20 | 19 | 0.05 |
| 5 | 1/6 | 20 | 14 | 1.8 |
| 6 | 1/6 | 20 | 17 | 0.45 |
| | | | **Total:** | **6.2** |

# Use a Chi-squared table to convert this to a probability

- Degrees of freedom = # bins - 1

Chi-Squared Table
Probability of rejecting the null hypothesis

| df | 99.5% | 99% | 97.5% | 95% | 90% | 10% | 5% | 2.5% | 1% | 0.5% |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7.88 | 6.64 | 5.02 | 3.84 | 2.71 | 0.02 | 0 | 0 | 0 | 0 |
| 2 | 10.6 | 9.21 | 7.38 | 5.99 | 4.61 | 0.21 | 0.1 | 0.05 | 0.02 | 0.01 |
| 3 | 12.84 | 11.35 | 9.35 | 7.82 | 6.25 | 0.58 | 0.35 | 0.22 | 0.12 | 0.07 |
| 4 | 14.86 | 13.28 | 11.14 | 9.49 | 7.78 | 1.06 | 0.71 | 0.48 | 0.3 | 0.21 |
| 5 | 16.75 | 15.09 | 12.83 | 11.07 | 9.24 | 1.61 | 1.15 | 0.83 | 0.55 | 0.41 |

Chi-Squared table. With 5 degrees of freedom (df) and $\chi^2 = 6.2$, the probability is between 90% and 10%

# StatTrek.com:

- Also works (easier too)

- Enter a value for degrees of freedom.
- Enter a value for one, and only one, of the remaining unshaded text boxes.
- Click the **Calculate** button to compute values for the other text boxes.

| | |
|---|---|
| Degrees of freedom | 5 |
| Chi-square critical value (CV) | 6.2 |
| $P(X^2 < 6.2)$ | 0.71 |
| $P(X^2 > 6.2)$ | 0.29 |

The probability the die is loaded is 0.71 (www.StatTrek.com)

# Repeat with 1,000,000 rolls of the dice:

```
RESULT = zeros(1,6);

n = 1e6;
p = 1/6;
for i=1:n
    d6 = ceil(rand*6);
    RESULT(d6) = RESULT(d6) + 1;
end

RESULT =   166220   166399   166933   167052   166500   166896

Chi2 = sum( (RESULT - n*p).^2) / (n*p)

Chi2 =    3.4257
```

This corresponds to a probability of 37%

- Not too large
- Not too small

## Example 2: Loaded Die

Suppose instead you had a loaded die:

- 90% of the time, the die is fair (all results have equal probability)
- 10% of the time, the result is always a 6.

Can you detect that the die is fair after 120 rolls?

Code:

```
n=1000;
for i=1:n
    if(rand < 0.1)
        d6 = 6;
    else
       d6 = ceil(rand*6);
    end
    RESULT(d6) = RESULT(d6) + 1;
end

RESULT =    24    18    10    17    21    30
```

Compute the Chi-Squared score:

$$\chi^2 = 11.5$$

| Die Roll (bin) | p theoretical probability | np expected frequency | N actual frequency | $\chi^2 = \left(\frac{(np-N)^2}{np}\right)$ |
|---|---|---|---|---|
| 1 | 1/6 | 20 | 24 | 0.8 |
| 2 | 1/6 | 20 | 18 | 0.2 |
| 3 | 1/6 | 20 | 10 | 5 |
| 4 | 1/6 | 20 | 17 | 0.45 |
| 5 | 1/6 | 20 | 21 | 0.05 |
| 6 | 1/6 | 20 | 30 | 5 |
| | | | **Total:** | 11.5 |

# Convert this to a probabity with a Chi-Squared table

- $\chi^2 = 11.5$
- 5 degrees of freedom (6 bins)
- p = 0.96 (from StatTrek)

Chi-Squared Table
Probability of rejecting the null hypothesis

| df | 99.5% | 99% | 97.5% | 95% | 90% | 10% | 5% | 2.5% | 1% | 0.5% |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7.88 | 6.64 | 5.02 | 3.84 | 2.71 | 0.02 | 0 | 0 | 0 | 0 |
| 2 | 10.6 | 9.21 | 7.38 | 5.99 | 4.61 | 0.21 | 0.1 | 0.05 | 0.02 | 0.01 |
| 3 | 12.84 | 11.35 | 9.35 | 7.82 | 6.25 | 0.58 | 0.35 | 0.22 | 0.12 | 0.07 |
| 4 | 14.86 | 13.28 | 11.14 | 9.49 | 7.78 | 1.06 | 0.71 | 0.48 | 0.3 | 0.21 |
| 5 | 16.75 | 15.09 | 12.83 | 11.07 | 9.24 | 1.61 | 1.15 | 0.83 | 0.55 | 0.41 |

Chi-Squared table. With 5 degrees of freedom (df) and $\chi^2$=11.5, the probability is about 95%

# Repeat for 1200 rolls:

```
RESULT = zeros(1,6);
N = 1200;
p = 1/6;

for i=1:N
    if(rand < 0.1)
        d6 = 6;
    else
       d6 = ceil(rand*6);
    end
    RESULT(d6) = RESULT(d6) + 1;
end

RESULT =   186   180   160   173   197   304

Chi2 = sum( (RESULT - N*p) .^ 2 ) / (N*p)

Chi2 =   68.7500
```

# With 1200 rolls, I'm > 99.95% certain this die is loaded

- and I'm probably broke after 1200 rolls of a loaded die

# Example 3: How loaded is too loaded?

- Load a die
- p(detection) = 5% after 120 rolls.

Solution: $\chi^2$ score = 11.1

- Assume x too many 6's rolled gives $\chi^2 = 11.2$
- Assume same number of rolls for all other numbers

- You can get away with 13.84 extra sixes
- p(loading) = 13.84 / 120 = 11.5%

| Die Roll (bin) | p theoretical probability | np expected frequency | N actual frequency | $\chi^2 = \left(\frac{(np-N)^2}{np}\right)$ |
|---|---|---|---|---|
| 1 | 1/6 | 20 | 20 - x/5 | $\left(\frac{(x/5)^2}{20}\right)$ |
| 2 | 1/6 | 20 | 20 - x/5 | $\left(\frac{(x/5)^2}{20}\right)$ |
| 3 | 1/6 | 20 | 20 - x/5 | $\left(\frac{(x/5)^2}{20}\right)$ |
| 4 | 1/6 | 20 | 20 - x/5 | $\left(\frac{(x/5)^2}{20}\right)$ |
| 5 | 1/6 | 20 | 20 - x/5 | $\left(\frac{(x/5)^2}{20}\right)$ |
| 6 | 1/6 | 20 | 20 + x | $\left(\frac{x^2}{20}\right)$ |
| | | | **Total:** | $\left(\frac{1.2x^2}{20}\right) = 11.5$ |

# Example 4: Fudging Data

Chi-Squared tests can also detect if data was fudged.

- Large $\chi^2$ means the data doesn't match the assumed distribution
- Small $\chi^2$ means the data fits the assumed discription too well (data was forged)
  - *It's possible but unlikely to get such good data*
- 

Chi-Squared Table

Probability of rejecting the null hypothesis

| df | 99.5% | 99% | 97.5% | 95% | 90% | 10% | 5% | 2.5% | 1% | 0.5% |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7.88 | 6.64 | 5.02 | 3.84 | 2.71 | 0.02 | 0 | 0 | 0 | 0 |
| 2 | 10.6 | 9.21 | 7.38 | 5.99 | 4.61 | 0.21 | 0.1 | 0.05 | 0.02 | 0.01 |
| 3 | 12.84 | 11.35 | 9.35 | 7.82 | 6.25 | 0.58 | 0.35 | 0.22 | 0.12 | 0.07 |
| 4 | 14.86 | 13.28 | 11.14 | 9.49 | 7.78 | 1.06 | 0.71 | 0.48 | 0.3 | 0.21 |
| 5 | 16.75 | 15.09 | 12.83 | 11.07 | 9.24 | 1.61 | 1.15 | 0.83 | 0.55 | 0.41 |

For example, suppose instead of rolling a fair die 1000 times I

- Only roll a die 100 times, and then
- Add 150 to the sum of each die roll so that it *looks* like I rolled the dice 1000 times.

You can spot this fake data by the data fitting the null hypothesis *too* well.

In Matlab:

```
RESULT = 150 * ones(1,6);

for i=1:100
    d6 = ceil(rand*6);
    RESULT(d6) = RESULT(d6) + 1;
end

RESULT

sum( ( (RESULT - 166.666).^2) / 166.666)
```

# Determine the $\chi^2$ score

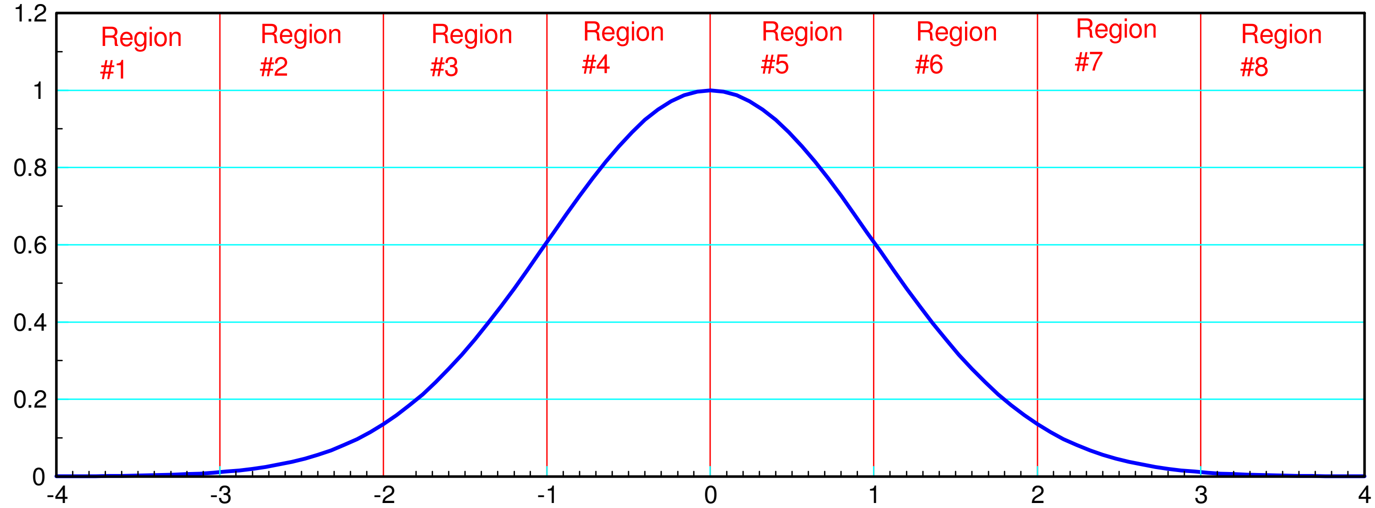
- $\chi^2 = 0.44$
- p = 0.5% (from StatTrek)
- The odds against getting such good data are 200 : 1 against.
- Most likely the data was faked.

| Die Roll (bin) | p theoretical probability | np expected frequency | N actual frequency | $\chi^2 = \left(\frac{(np-N)^2}{np}\right)$ |
|---|---|---|---|---|
| 1 | 1/6 | 166.67 | 168 | 0.01 |
| 2 | 1/6 | 166.67 | 168 | 0.01 |
| 3 | 1/6 | 166.67 | 159 | 0.35 |
| 4 | 1/6 | 166.67 | 169 | 0.03 |
| 5 | 1/6 | 166.67 | 169 | 0.03 |
| 6 | 1/6 | 166.67 | 167 | 0 |
| | | | **Total:** | 0.44 |

# Chi-Squared with Continuous Distributions

Also works with continuous distributions

- Split the continuous variable into N distinct regions (many ways to do this)
- Calculate the probability that any given data point will fall into each region (p)
- Calculate the expected number of observations you should have in each region (np),
- Compare to the actual frequency (N), i.e. calculate the $\chi^2$ score, then
- Convert the chi-squared score into a probability.



## Example:

```
X = sum( rand(12,1) ) - 6
```
$\approx N(0, 1)$

## Is this a standard Normal distribution?

- No - I can see the code

## Can you detect that it is not a standard normal variable?

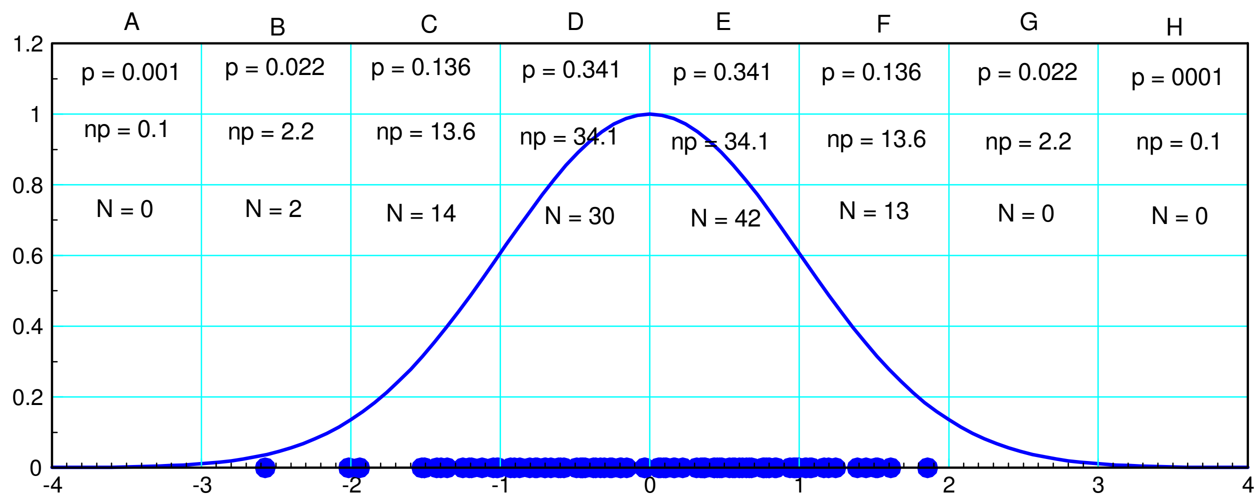
- harder.
- Use a $\chi^2$ table

# Example: Generate 100 random numbers

```
X = [];

for i=1:100
   X = [X ; sum( rand(12,1) ) - 6];
   end
```

- Split the X axis into 8 regions (A..H) (this is somewhat arbitrary).
- Compute the probability of each region (p) and the expected frequency (np)
- Count how many times X fell into each region (N)
- From this, create a Chi-Squared table



# Compute the $\chi^2$ score

- $\chi^2 = 4.7906$
- p = 0.31 (from StatTrek)
- Can't tell with only 100 data points

| Region (bin) | p | np | N | $\chi^2 = \left(\frac{(np-N)^2}{np}\right)$ |
|---|---|---|---|---|
| A | 0.001 | 0.1 | 0 | 0.1 |
| B | 0.022 | 2.2 | 2 | 0.0182 |
| C | 0.138 | 13.8 | 14 | 0.0029 |
| D | 0.341 | 34.1 | 30 | 0.493 |
| E | 0.341 | 34.1 | 42 | 1.8302 |
| F | 0.138 | 13.8 | 13 | 0.0464 |
| G | 0.022 | 2.2 | 0 | 2.2 |
| H | 0.001 | 0.1 | 0 | 0.1 |
| | | | **Total:** | **4.7906** |

## Repeat with 100,000 random numbers.

- $\chi^2 = 36.9886$
- p > 0.9995 (from StatTrek)
- With enough data I can tell that this isn't really a standard normal distribution
- The information is in the tails.

| Region (bin) | p | np | N | $\chi^2 = \left(\frac{(np-N)^2}{np}\right)$ |
|---|---|---|---|---|
| A | 0.0013 | 132 | 97 | 9.2803 |
| B | 0.0214 | 2,140 | 2,085 | 1.4136 |
| C | 0.1359 | 13,591 | 13,751 | 1.8836 |
| D | 0.3413 | 34,134 | 34,067 | 0.1315 |
| E | 0.3413 | 34,134 | 33,845 | 2.4469 |
| F | 0.1359 | 13,591 | 13,895 | 6.7998 |
| G | 0.0214 | 2,140 | 2,168 | 0.3664 |
| H | 0.0013 | 132 | 88 | 14.6667 |
| | | | **Total:** | **36.9886** |