# Sta 111 - Summer II 2017
# Probability and Statistical Inference

## 15. Chi-square test of goodness of fit

Lu Wang

Duke University, Department of Statistical Science

July 25, 2017

# Outline

1. Goodness of fit

2. The chi-square test statistic

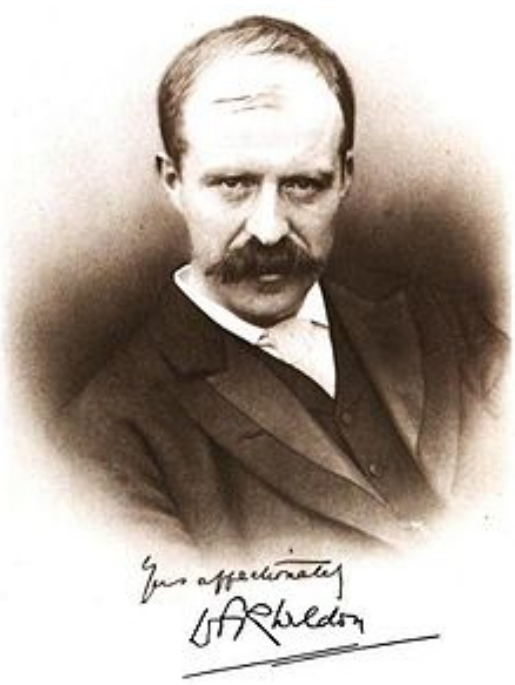
3. The chi-square distribution and finding areas
    1. Degrees of freedom for a goodness of fit test
    2. P-value for a chi-square test

4. Conditions for the chi-square test

5. Summary

# Weldon's dice

- Walter Frank Raphael Weldon (1860 - 1906), was an English evolutionary biologist and a founder of biometry. He was the joint founding editor of Biometrika, with Francis Galton and Karl Pearson.
- In 1894, he rolled 12 dice 26,306 times, and recorded the number of 5s or 6s (which he considered to be a success).



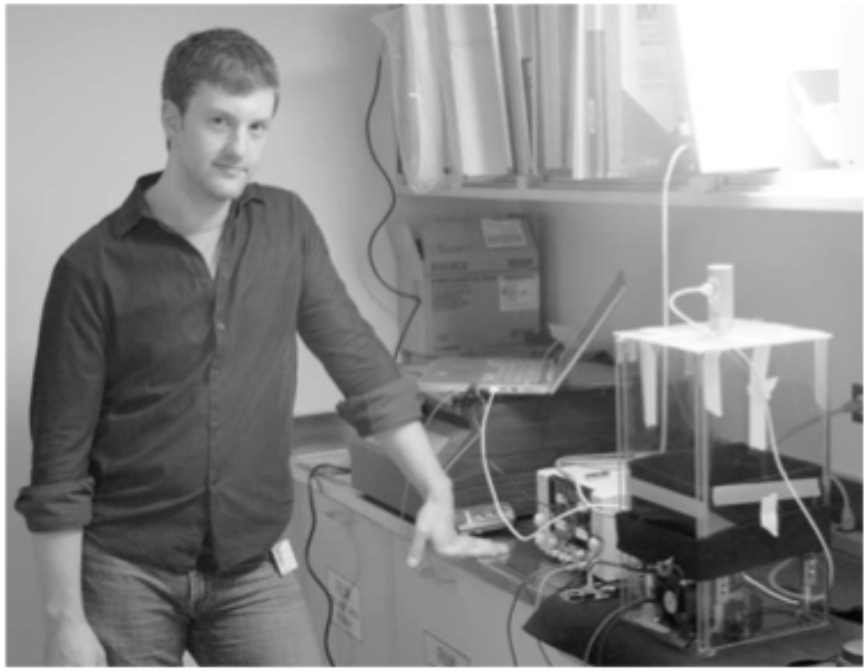
  - It was observed that 5s or 6s occurred more often than expected, and Pearson hypothesized that this was probably due to the construction of the dice. Most inexpensive dice have hollowed-out pips, and since opposite sides add to 7, the face with 6 pips is lighter than its opposing face, which has only 1 pip.

## Labby's dice

- In 2009, Zacariah Labby (U of Chicago), repeated Weldon's experiment using a homemade dice-throwing, pip counting machine.

  *http://www.youtube.com/watch?v=95EErdouO2w*

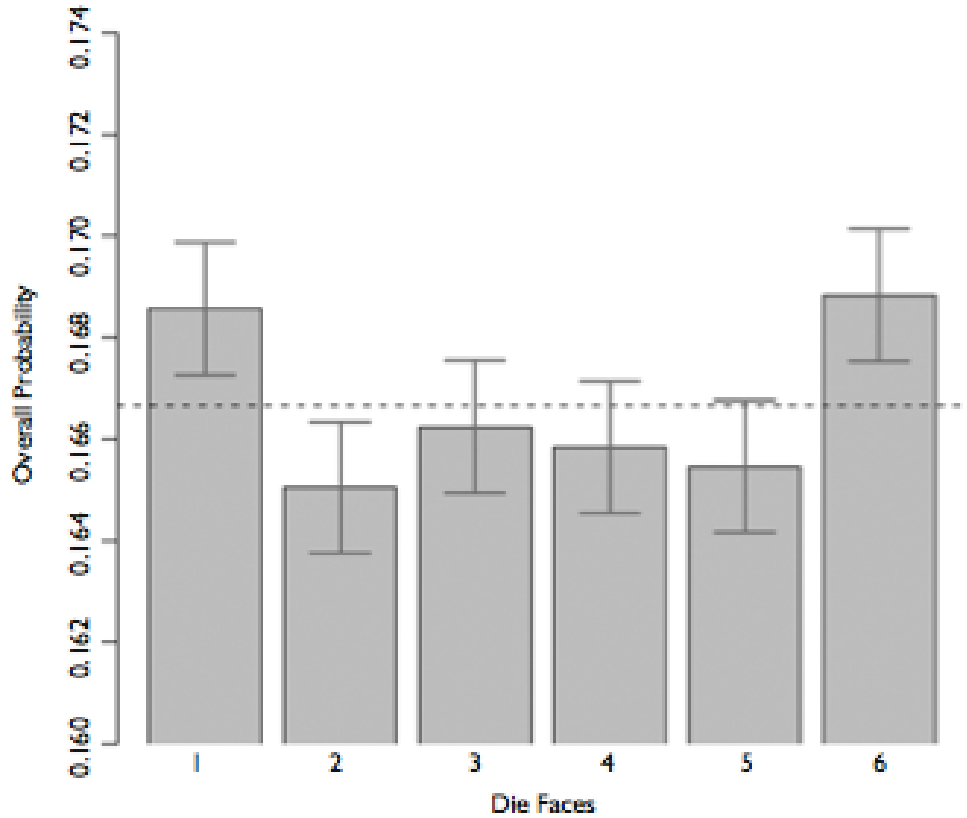
- The rolling-imaging process took about 20 seconds per roll.



  - Each day there were ∼150 images to process manually.
  - At this rate Weldon's experiment was repeated in a little more than six full days.
  - Recommended reading: *http://galton.uchicago.edu/about/docs/labby09dice.pdf*

# Labby's dice (cont.)

- Labby did not actually observe the same phenomenon that Weldon observed (higher frequency of 5s and 6s).
- Automation allowed Labby to collect more data than Weldon did in 1894, instead of recording “successes” and “failures”, Labby recorded the individual number of pips on each die.



# Expected counts

Labby rolled 12 dice 26,306 times. If each side is equally likely to come up, how many 1s, 2s, $\cdots$, 6s would he expect to have observed?

(a) $\frac{1}{6}$

(b) $\frac{12}{6}$

(c) $\frac{26{,}306}{6}$

(d) $\frac{12 \times 26{,}306}{6}$

## Summarizing Labby's results

The table below shows the observed and expected counts from Labby's experiment.

| Outcome | Observed | Expected |
|---|---|---|
| 1 | 53,222 | 52,612 |
| 2 | 52,118 | 52,612 |
| 3 | 52,465 | 52,612 |
| 4 | 52,338 | 52,612 |
| 5 | 52,244 | 52,612 |
| 6 | 53,285 | 52,612 |
| Total | 315,672 | 315,672 |

At a first glance, does there appear to be an inconsistency between the observed and expected counts?

## Setting the hypotheses

Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

$H_0$: There is no inconsistency between the observed and the expected counts, i.e. *each side is equally likely to come up.* Variations in observed counts reflect natural sampling fluctuation.

$H_A$: There is an inconsistency between the observed and the expected counts, i.e. *there is a bias in which side comes up on the roll of a die.*

## Goodness of fit

- To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts.

- Large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis.

- This is called a *goodness of fit* test since we're evaluating *how well the observed data fit the expected distribution.*

## Anatomy of a test statistic

- The general form of a test statistic is

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

- This construction is based on
  1. identifying the difference between a point estimate and an expected value if the null hypothesis was true, and
  2. standardizing that difference using the standard error of the point estimate.

These two ideas will help in the construction of an appropriate test statistic for count data.

When dealing with *counts* and investigating how far the observed counts are from the expected counts, we use a new test statistic called the *chi-square ($\chi^2$) statistic*.

## $\chi^2$ statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- $O_i$ is the observed count of category $i$
- $E_i$ is the expected count of category $i$
- $k$ is the total number of categories

## Why square?

Squaring the difference between the observed and the expected outcome does two things:

- Any standardized difference that is squared will now be positive.
- Differences that already looked unusual will become much larger after being squared.

Where have we seen this before?

**Important points:**

- Use **counts** (not **proportions**) in the calculation of the $\chi^2$ statistic, even though we're truly interested in the proportions for inference
- Expected counts are calculated assuming the null hypothesis is true

# Calculating the chi-square statistic

| Outcome | Observed | Expected | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|
| 1 | 53,222 | 52,612 | $\frac{(53{,}222-52{,}612)^2}{52{,}612} = 7.07$ |
| 2 | 52,118 | 52,612 | $\frac{(52{,}118-52{,}612)^2}{52{,}612} = 4.64$ |
| 3 | 52,465 | 52,612 | $\frac{(52{,}465-52{,}612)^2}{52{,}612} = 0.41$ |
| 4 | 52,338 | 52,612 | $\frac{(52{,}338-52{,}612)^2}{52{,}612} = 1.43$ |
| 5 | 52,244 | 52,612 | $\frac{(52{,}244-52{,}612)^2}{52{,}612} = 2.57$ |
| 6 | 53,285 | 52,612 | $\frac{(53{,}285-52{,}612)^2}{52{,}612} = 8.61$ |
| Total | 315,672 | 315,672 | 24.73 |

## The chi-square distribution

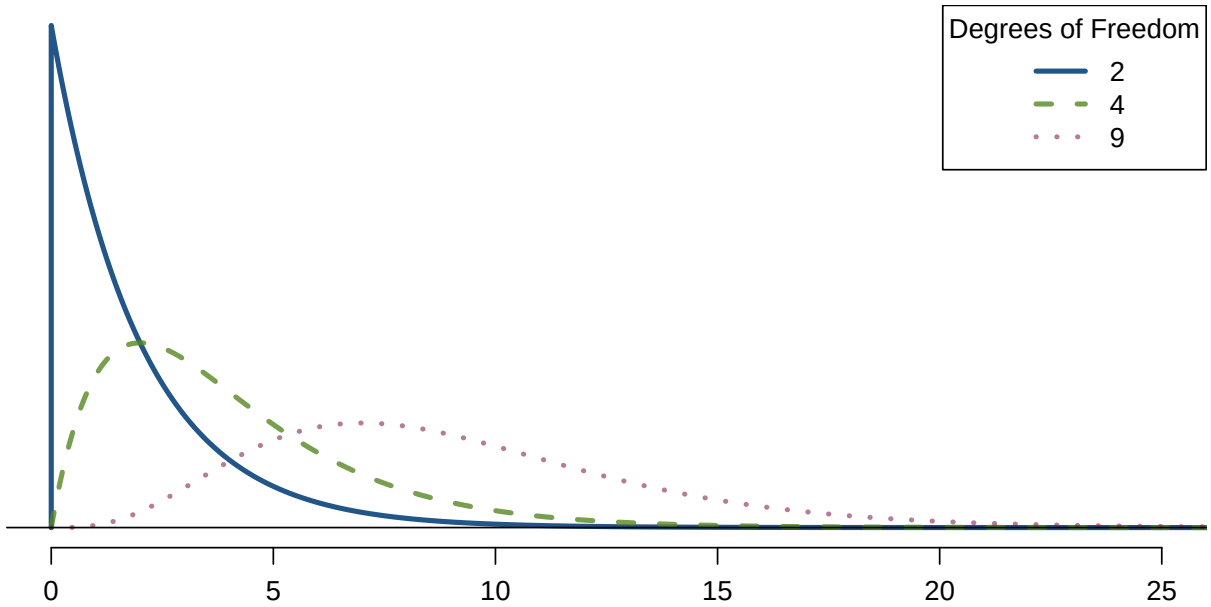
- In order to determine if the $\chi^2$ statistic we calculated is considered unusually high or not, we need to first describe its distribution.
- The chi-square distribution has just one parameter called *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.

*Recap:* so far we've seen three other continuous distributions:

- normal distribution: unimodal and symmetric with two parameters: mean and standard deviation
- T distribution: unimodal and symmetric with one parameter: degrees of freedom
- F distribution: unimodal and right skewed with two parameters: degrees of freedom for numerator (between group variance) and denominator (within group variance)

## A $\chi^2$ random variable is always positive and right skewed

### Which of the following is false?




As the df increases,

(a) the center of the $\chi^2$ distribution increases as well

(b) the variability of the $\chi^2$ distribution increases as well

(c) the shape of the $\chi^2$ distribution becomes more skewed (less like a normal)

## Degrees of freedom for a goodness of fit test

- When conducting a goodness of fit test to evaluate how well the observed data follow an expected distribution, the degrees of freedom are calculated as the number of categories ($k$) minus 1.

$$df = k - 1$$

- For dice outcomes, $k = 6$, therefore

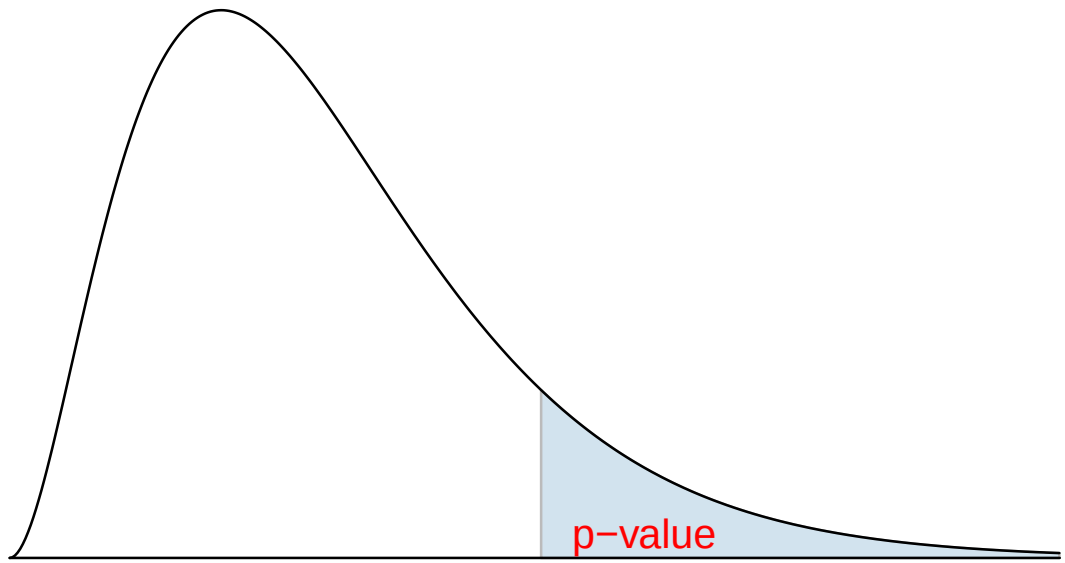
$$df = 6 - 1 = 5$$

**Back to Labby's dice**

- We had calculated a test statistic of $\chi^2 = 24.67$ with $df = 5$.
- All we need is the p-value and we can make a decision on the hypotheses.

## Finding areas under the chi-square curve

- The p-value for a chi-square test is defined as the tail area *above* the calculated test statistic.
- This is because the test statistic is always positive, and a higher test statistic means a stronger deviation from the null hypothesis.



## Conclusion of the hypothesis test - Labby's dice

p-value = $P(\chi^2_{df=5} > 24.67) \approx 0.0002$

```
> pchisq(q=24.67, df=5, lower.tail = FALSE)
[1] 0.0001613338
```

At 5% significance level, what is the conclusion of the hypothesis test?

(a) Reject $H_0$, the data provide convincing evidence that the dice are fair.

(b) Reject $H_0$, the data provide convincing evidence that the dice are biased.

(c) Fail to reject $H_0$, the data provide convincing evidence that the dice are fair.

(d) Fail to reject $H_0$, the data provide convincing evidence that the dice are biased.

## Turns out...

- The faces with one and six pips have consistently higher chance to come up than the other faces.
- Pearson's claim that 5s and 6s appear more often due to the carved-out pips is not supported by these data.
- Dice used in casinos have flush faces, where the pips are filled in with a plastic of the same density as the surrounding material and are precisely balanced.

## Conditions for the chi-square test

1. *Independence:* Each case that contributes a count to the table must be independent of all the other cases in the table.

2. *Sample size:* Each particular category must have at least 5 *expected* cases.

3. *df* $>$ *1:* Degrees of freedom must be greater than 1 (at least 3 categories).

- Failing to check conditions may affect the test's error rates.
- **If sample size related conditions are not met:** Simulation based inference (randomization for HT / bootstrapping for CI, when appropriate)

## Practice - 2009 Iran Election

There was lots of talk of election fraud in the 2009 Iran election. We'll compare the data from a poll conducted before the election (observed data) to the reported votes in the election to see if the two follow the same distribution.

| Candidate | Observed # of voters in poll | Reported % of votes in election |
|---|---|---|
| (1) Ahmedinajad | 338 | 63.29% |
| (2) Mousavi | 136 | 34.10% |
| (3) Minor candidates | 30 | 2.61% |
| Total | 504 | 100% |
| | ↓ | ↓ |
| | *observed* | *expected distribution* |

## Hypotheses

What are the hypotheses for testing if the distributions of reported and polled votes are different?

$H_0$: The observed counts from the poll follow the same distribution as the reported votes.

$H_A$: The observed counts from the poll do not follow the same distribution as the reported votes.

## Calculation of the test statistic

First check conditions for a chi-square test!

| Candidate | Observed # of voters in poll | Reported % of votes in election | Expected # of votes in poll |
|---|---|---|---|
| (1) Ahmedinajad | 338 | 63.29% | $504 \times 0.6329 = 319$ |
| (2) Mousavi | 136 | 34.10% | $504 \times 0.3410 = 172$ |
| (3) Minor candidates | 30 | 2.61% | $504 \times 0.0261 = 13$ |
| Total | 504 | 100% | 504 |

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(338 - 319)^2}{319} = 1.13$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(136 - 172)^2}{172} = 7.53$$

$$\frac{(O_3 - E_3)^2}{E_3} = \frac{(30 - 13)^2}{13} = 22.23$$

$$\chi^2_{df=3-1=2} = 30.89$$

## Conclusion

```
> pchisq(q=30.89, df=2, lower.tail = FALSE)
[1] 1.960296e-07
```

Based on these calculations what is the conclusion of the hypothesis test?

(a) p-value is low, $H_0$ is rejected. The observed counts from the poll do not follow the same distribution as the reported votes.

(b) p-value is high, $H_0$ is not rejected. The observed counts from the poll follow the same distribution as the reported votes.

(c) p-value is low, $H_0$ is rejected. The observed counts from the poll follow the same distribution as the reported votes

(d) p-value is low, $H_0$ is not rejected. The observed counts from the poll do not follow the same distribution as the reported votes.

**If sample size related conditions are met:**

- Categorical data with 2 levels $\rightarrow$ Z
  - one variable: Z HT / CI for a single proportion
  - two variables: Z HT / CI comparing two proportions
- Categorical data with more than 2 levels $\rightarrow \chi^2$
  - one variable: $\chi^2$ *test of goodness of fit*, no CI
  - two variables: $\chi^2$ *test of independence*, no CI

**If sample size related conditions are not met:** Simulation based inference (randomization for HT / bootstrapping for CI, when appropriate)