Inference for Two Way Tables

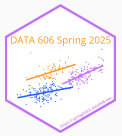
IS381 - Statistics and Probability with R

Jason Bryer, Ph.D.

Week 10

Weldon's dice

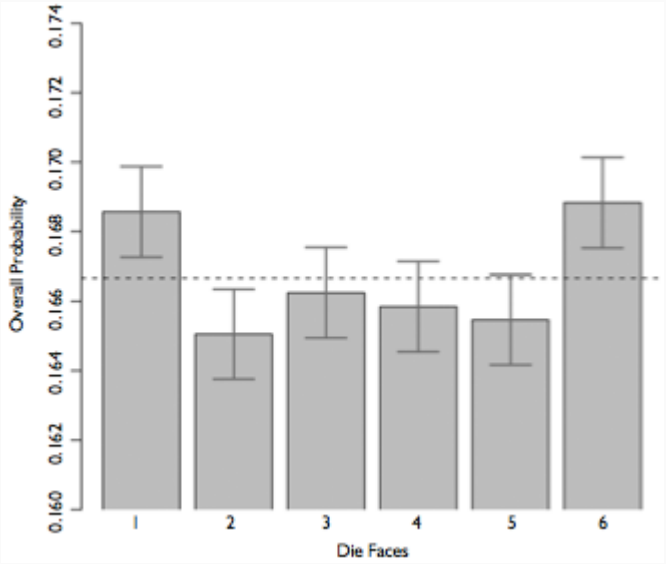
- Walter Frank Raphael Weldon (1860 - 1906), was an English evolutionary biologist and a founder of biometry. He was the joint founding editor of Biometrika, with Francis Galton and Karl Pearson.
- In 1894, he rolled 12 dice 26,306 times, and recorded the number of 5s or 6s (which he considered to be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

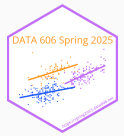
- The rolling-imaging process took about 20 seconds per roll.
  - Each day there were ~150 images to process manually.
  - At this rate Weldon's experiment was repeated in a little more than six full days.



Summarizing Labby's results

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

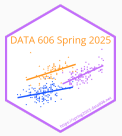
Outcome	Observed	Expected
1	53,222	52,612
2	52,118	52,612
3	52,465	52,612
4	52,338	52,612
5	52,244	52,612
6	53,285	52,612
Total	315,672	315,672



Setting the hypotheses

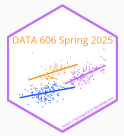
Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

- H_0: There is no inconsistency between the observed and the expected counts. The observed counts follow the same distribution as the expected counts.
- H_A: There is an inconsistency between the observed and the expected counts. The observed counts **do not** follow the same distribution as the expected counts. There is a bias in which side comes up on the roll of a die.



Evaluating the hypotheses

- 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

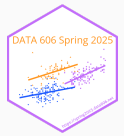
- The general form of a test statistic is:

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

- 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.



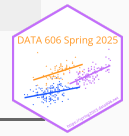
Chi-Squared

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 (χ^2) statistic.

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

where k = total number of cells

Outcome	Observed	Expected	$\frac{(O-E)^2}{E}$
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



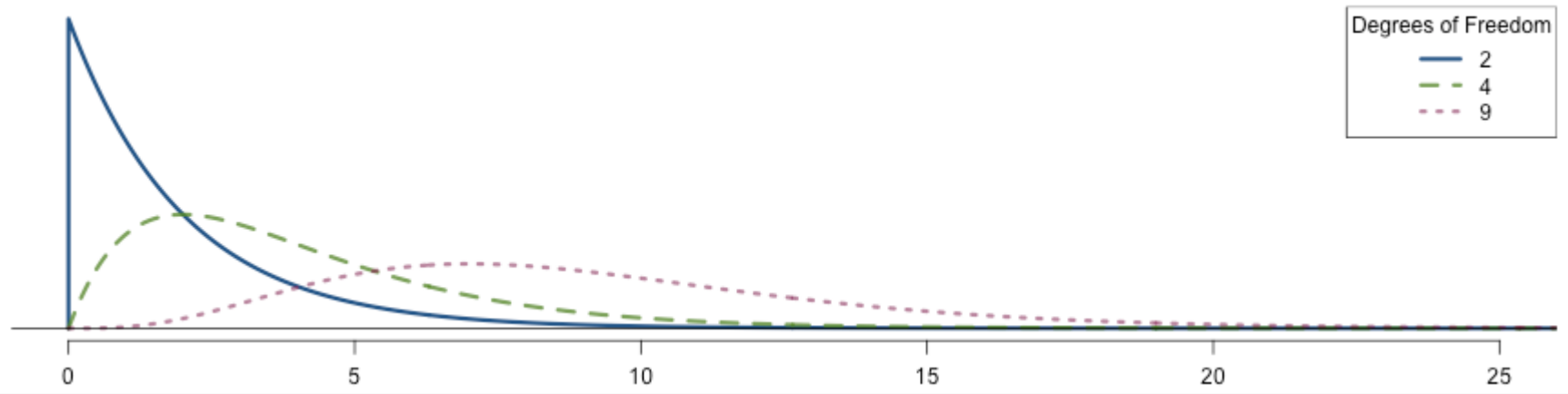
Chi-Squared Distribution

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.

In order to determine if the χ^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.

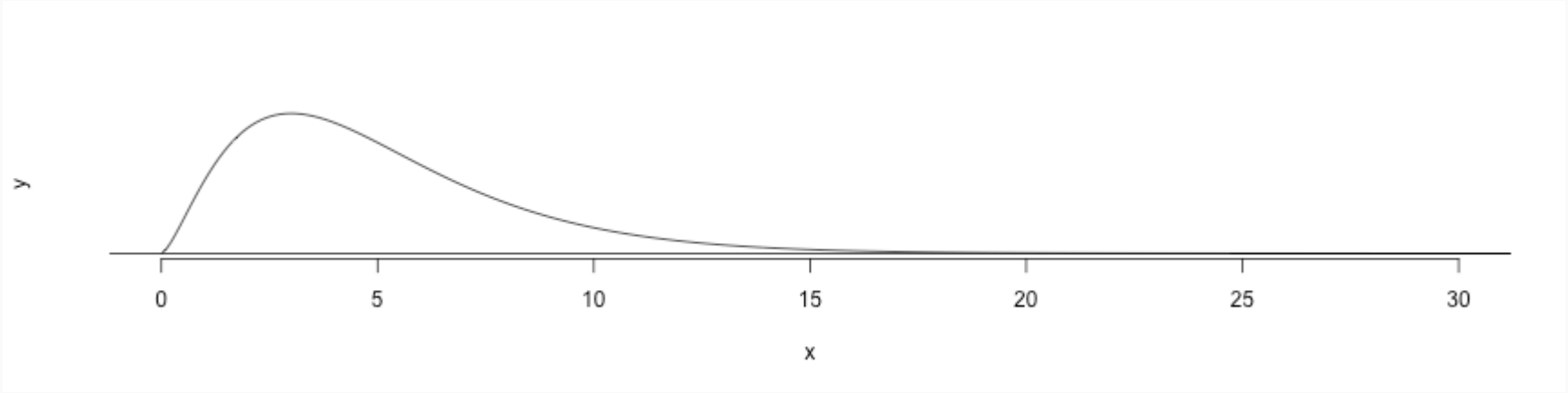


Degrees of freedom for a goodness of fit test

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 cells (k) minus 1.

$$df = k - 1$$

For dice outcomes, $k = 6$, therefore $df = 6 - 1 = 5$



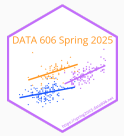
p-value = $P(\chi^2_{df=5} > 24.67)$ is less than 0.001

Turns out...

- The 1-6 axis is consistently shorter than the other two (2-5 and 3-4), thereby supporting the hypothesis that the faces with one and six pips are larger than the other faces.

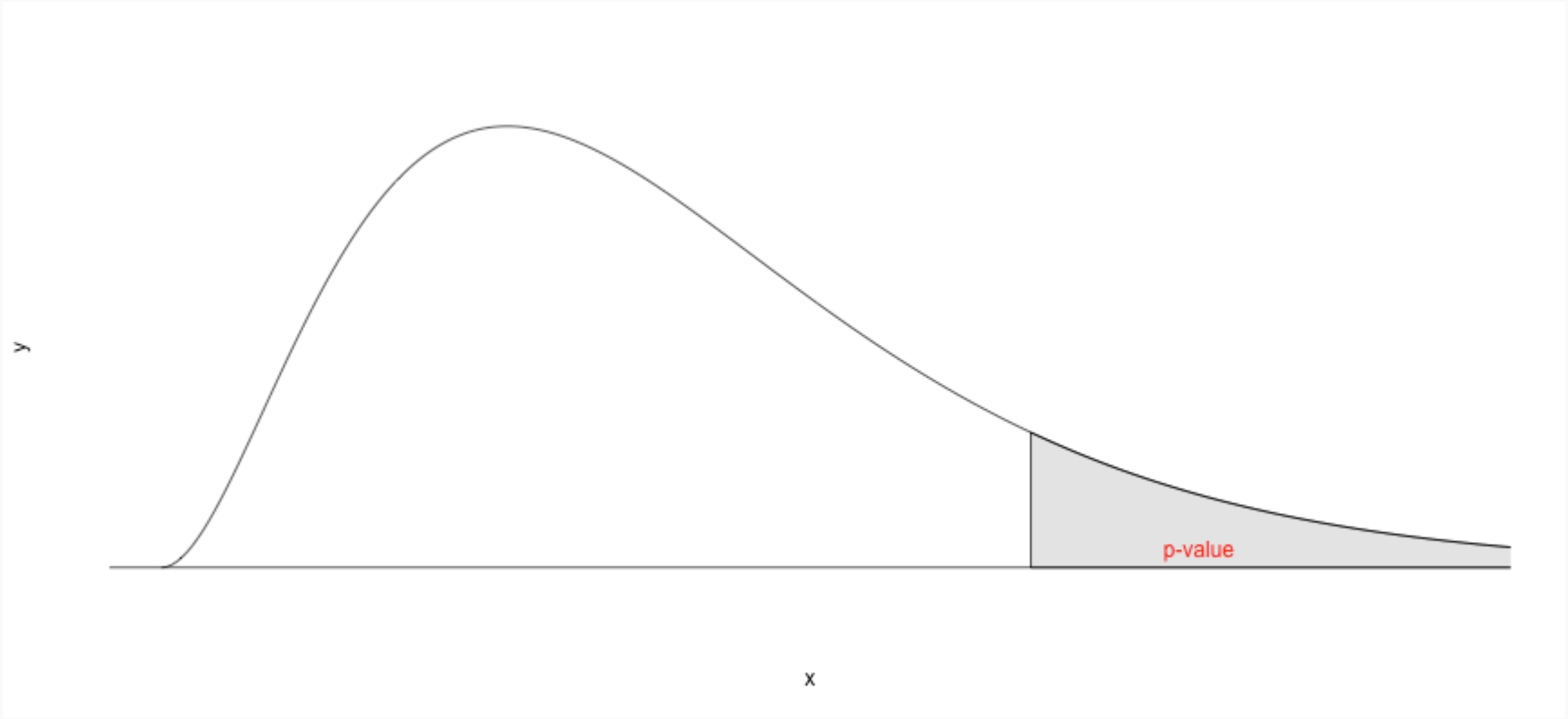
- Pearson's claim that 5s and 6s appear more often due to the carved-out pips is not supported by these data.

- 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.



Recap: p-value for a chi-square test

- 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.

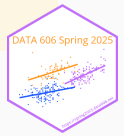


Independence Between Groups

Assume we have a population of 100,000 where groups A and B are independent with $p_A = .55$ and $p_B = .6$ and $n_A = 99,000$ (99% of the population) and $n_B = 1,000$ (1% of the population). We can sample from the population (that includes groups A and B) and from group B of sample sizes of 1,000 and 100, respectively. We can also calculate $\hat{p}$ for group A independent of B.

```
propA <- .55    # Proportion for group A
propB <- .6     # Proportion for group B
pop.n <- 100000 # Population size
sampleA.n <- 1000
sampleB.n <- 100
```

```
pop <- data.frame(
    group = c(rep('A', pop.n * 0.99),
              rep('B', pop.n * 0.01) ),
    response = c(
        sample(c(1,0),
               size = pop.n * 0.99,
               prob = c(propA, 1 - propA),
               replace = TRUE),
        sample(c(1,0),
               size = pop.n * 0.01,
               prob = c(propB, 1 - propB),
               replace = TRUE) )
)
sampA <- pop[sample(nrow(pop),
                    size = sampleA.n),]
sampB <- pop[sample(which(pop$group == 'B'),
                    size = sampleB.n),]
```



Independence Between Groups (cont.)

$\hat{p}$ for the population sample

```
mean(sampA$response)
```

```
## [1] 0.57
```

$\hat{p}$ for the population sample, excluding group B

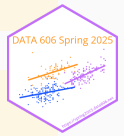
```
mean(sampA[sampA$group == 'A',]$response)
```

```
## [1] 0.5679758
```

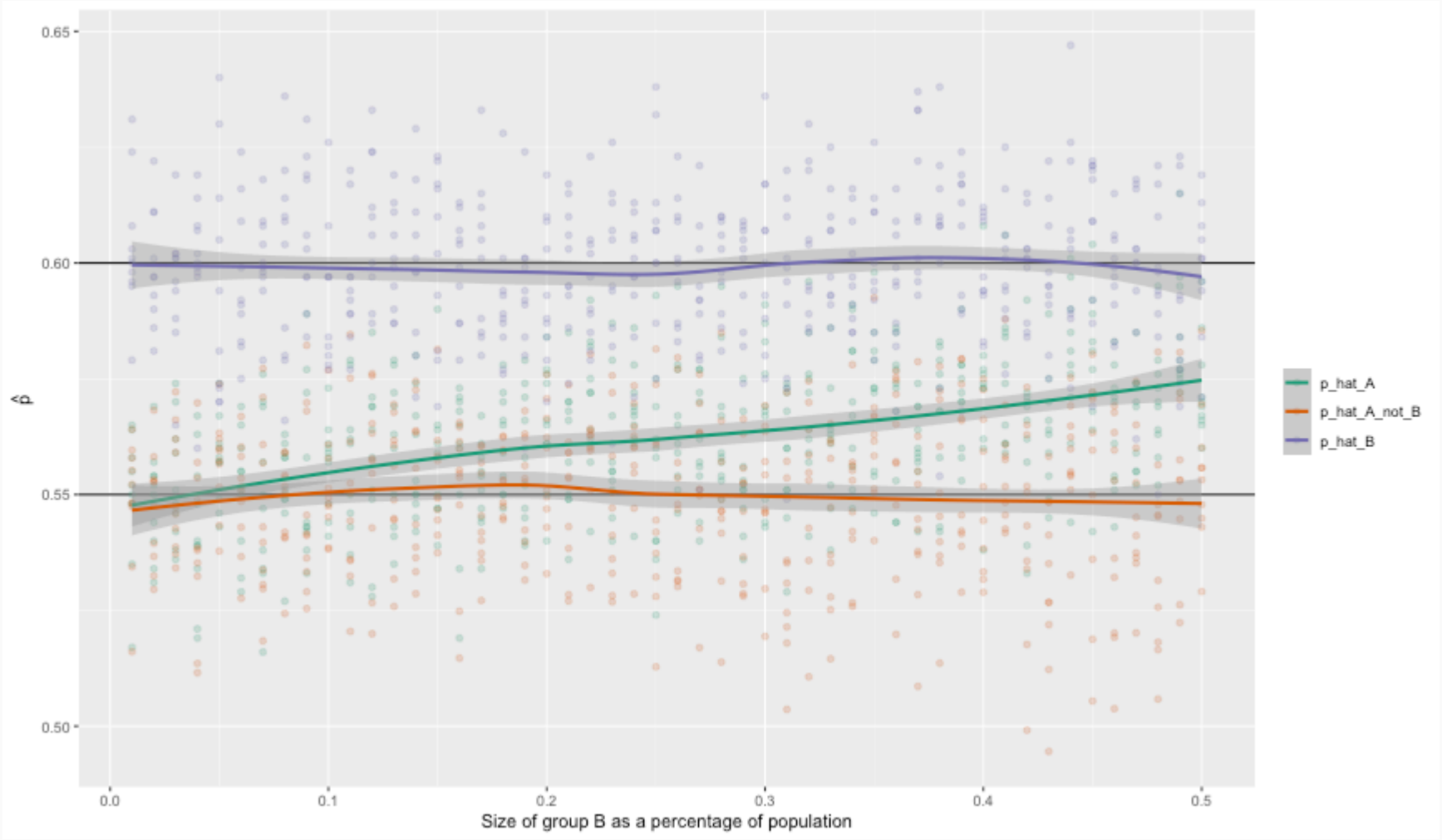
$\hat{p}$ for group B sample

```
mean(sampB$response)
```

```
## [1] 0.66
```



Independence Between Groups (cont.)



One Minute Paper

1. What was the most important thing you learned during this class?

2. What important question remains unanswered for you?

Scan the QR code or click here to complete the One Minute Paper