

**Welcome to Week #10!**

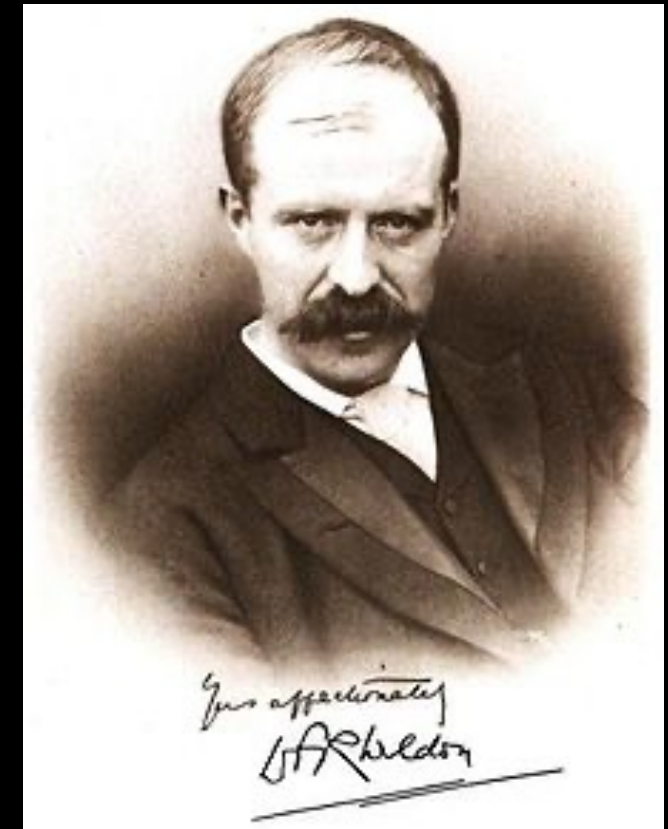
**Foundations for Inference - How well can we really know anything?**

**$\chi^2$  & ANOVA - For more complex datasets**

# $\chi^2$ - When and Why: Dice Example

Walter Frank Raphael Weldon (1860 - 1906), was an English evolutionary biologist and a founder of biometry (application of stats to bio data).

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.

# $\chi^2$ - When and Why: Dice Example

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

[www.youtube.com/watch?v=95EErdouO2w](http://www.youtube.com/watch?v=95EErdouO2w)

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

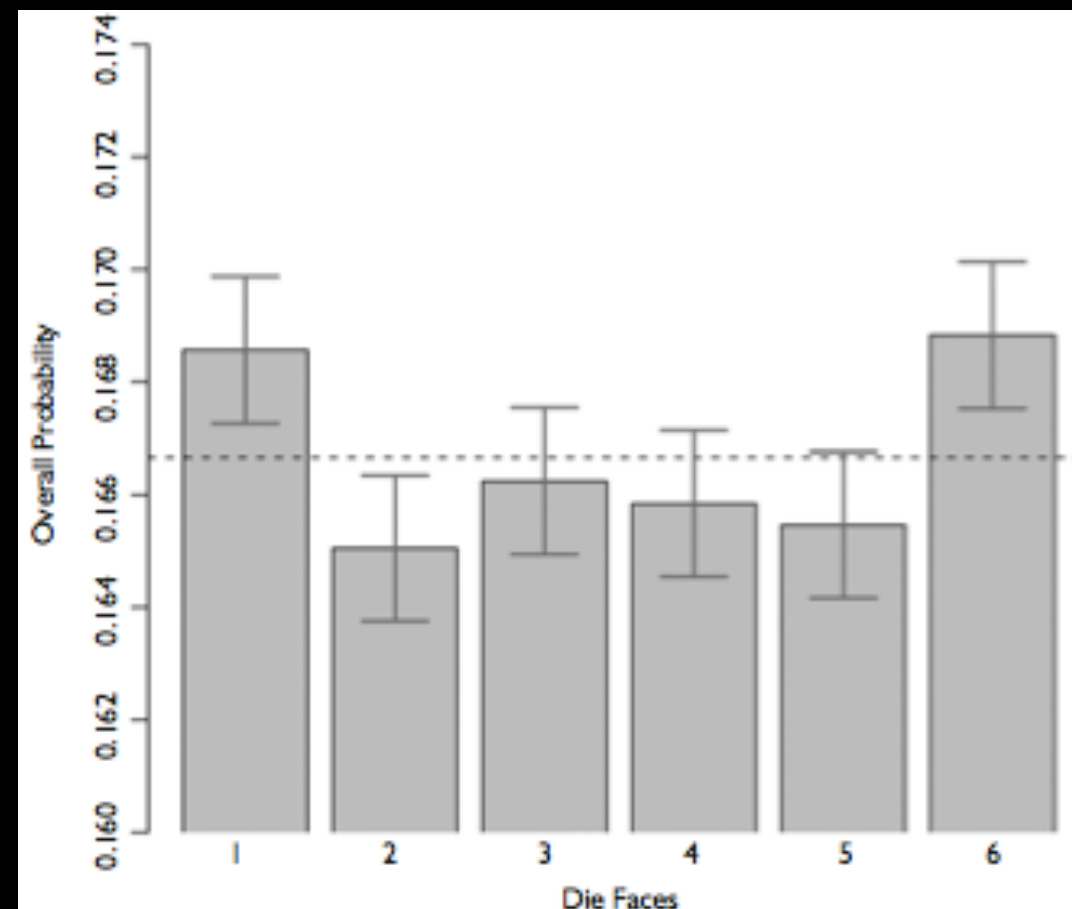
Recommended reading:

[galton.uchicago.edu/about/docs/labby09dice.pdf](http://galton.uchicago.edu/about/docs/labby09dice.pdf)

# $\chi^2$ - When and Why: Dice Example

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.



# $\chi^2$ - When and Why: Dice Example

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

(number of dice X number of rolls)/(number of sides)

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

$$= 12 \times 26,306 / 6 = 52,612$$

# 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.

# Chi-square statistic

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**.

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

$\chi^2$  statistic

$$\chi^2 = \sum_{i=1}^k \frac{(O - E)^2}{E} \quad \text{where } k = \text{total number of cells}$$

# Calculating the chi-square statistic

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

# Calculating the chi-square statistic

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	$\frac{(52,465-52,612)^2}{52,612} = 0.41$

## Why Squared?

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

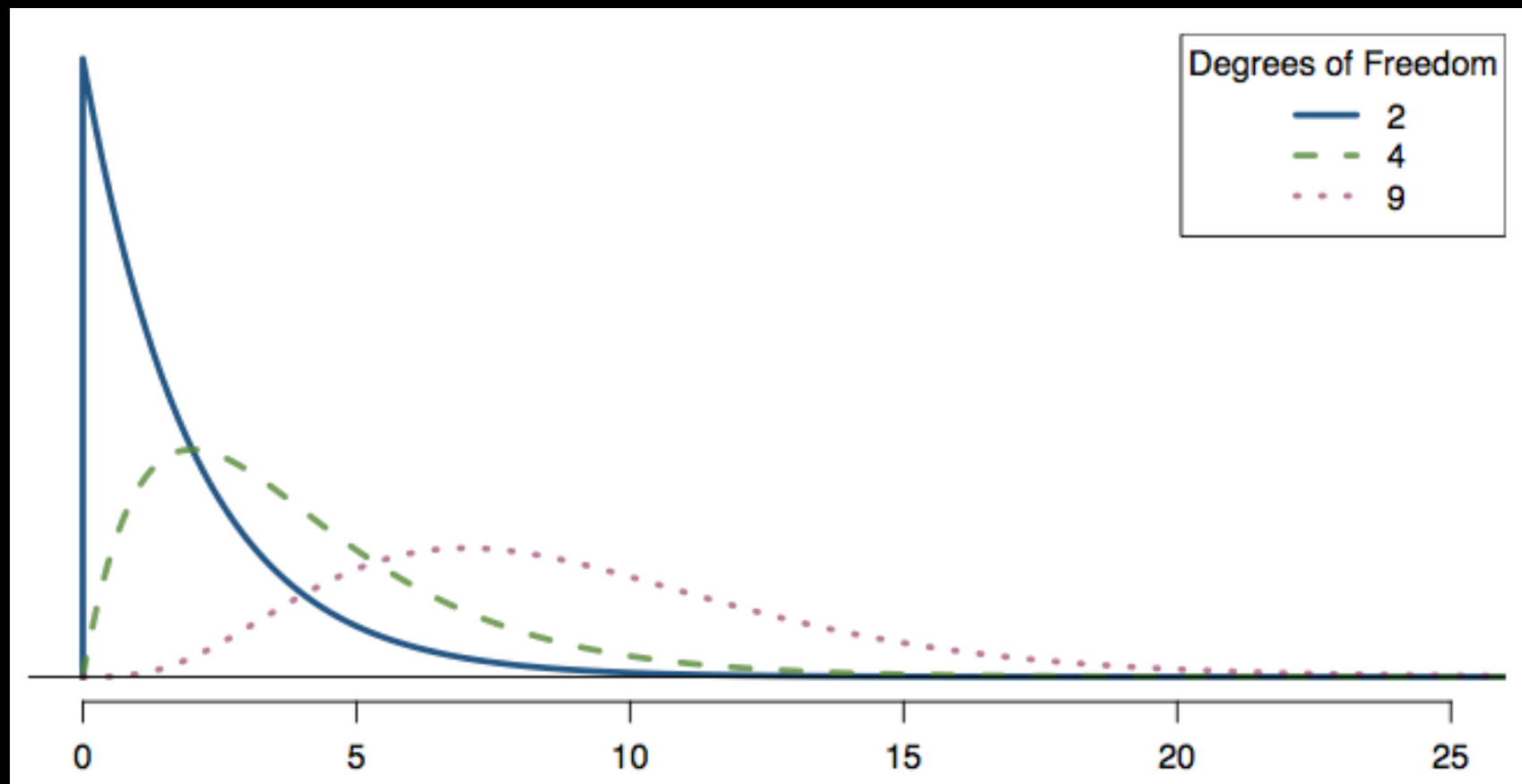
- Any standardized difference that is squared will now be positive.
- Differences that already looked unusual will become much larger after being squared.

Total	315,672	315,672	24.73
-------	---------	---------	-------

# The chi-square distribution

In order to determine if the  $\chi^2$  statistic we calculated (24.73) 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.



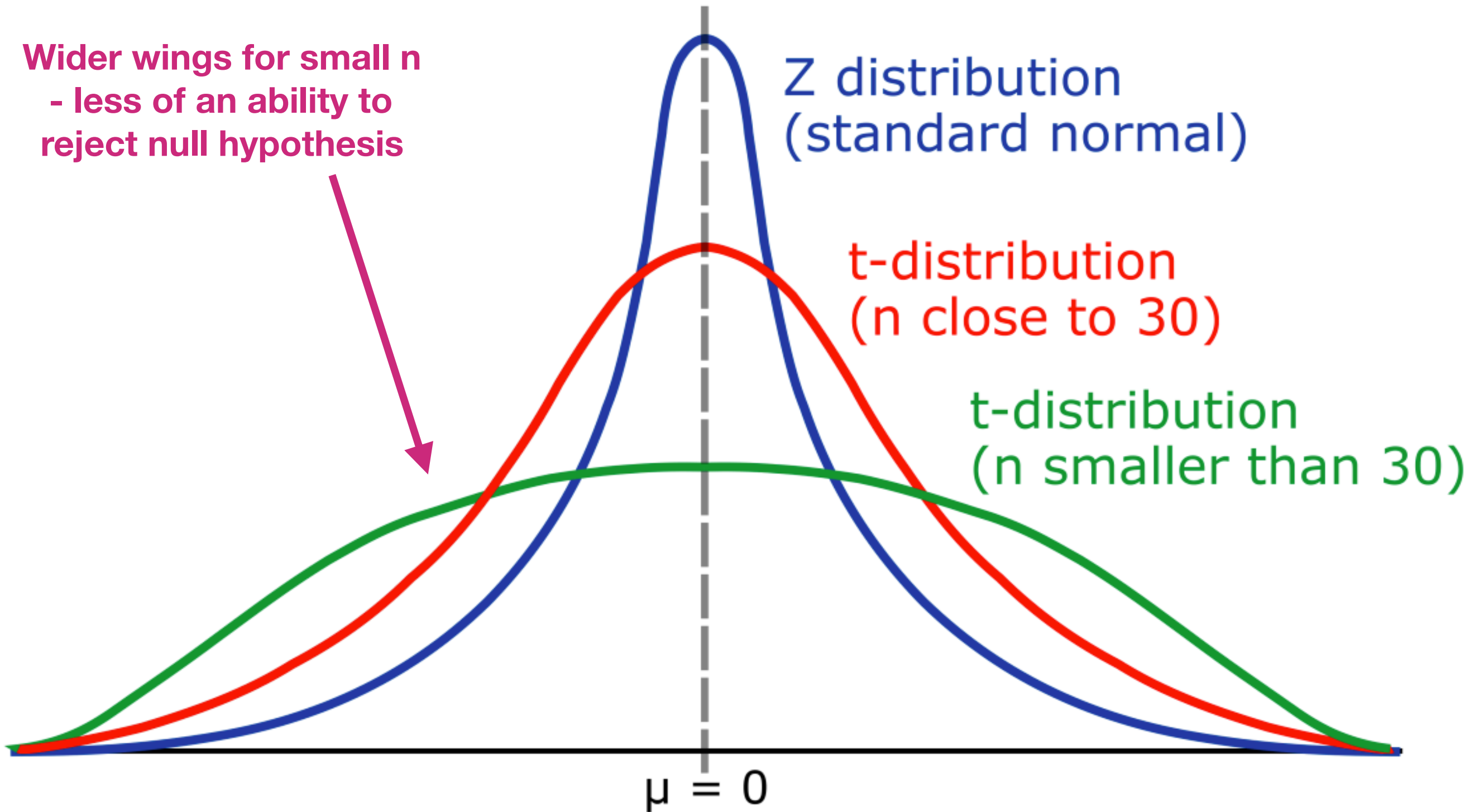
Wider wings for small n  
- less of an ability to  
reject null hypothesis

Z distribution  
(standard normal)

t-distribution  
(n close to 30)

t-distribution  
(n smaller than 30)

$\mu = 0$



Wider wings for small n  
- less of an ability to  
reject null hypothesis

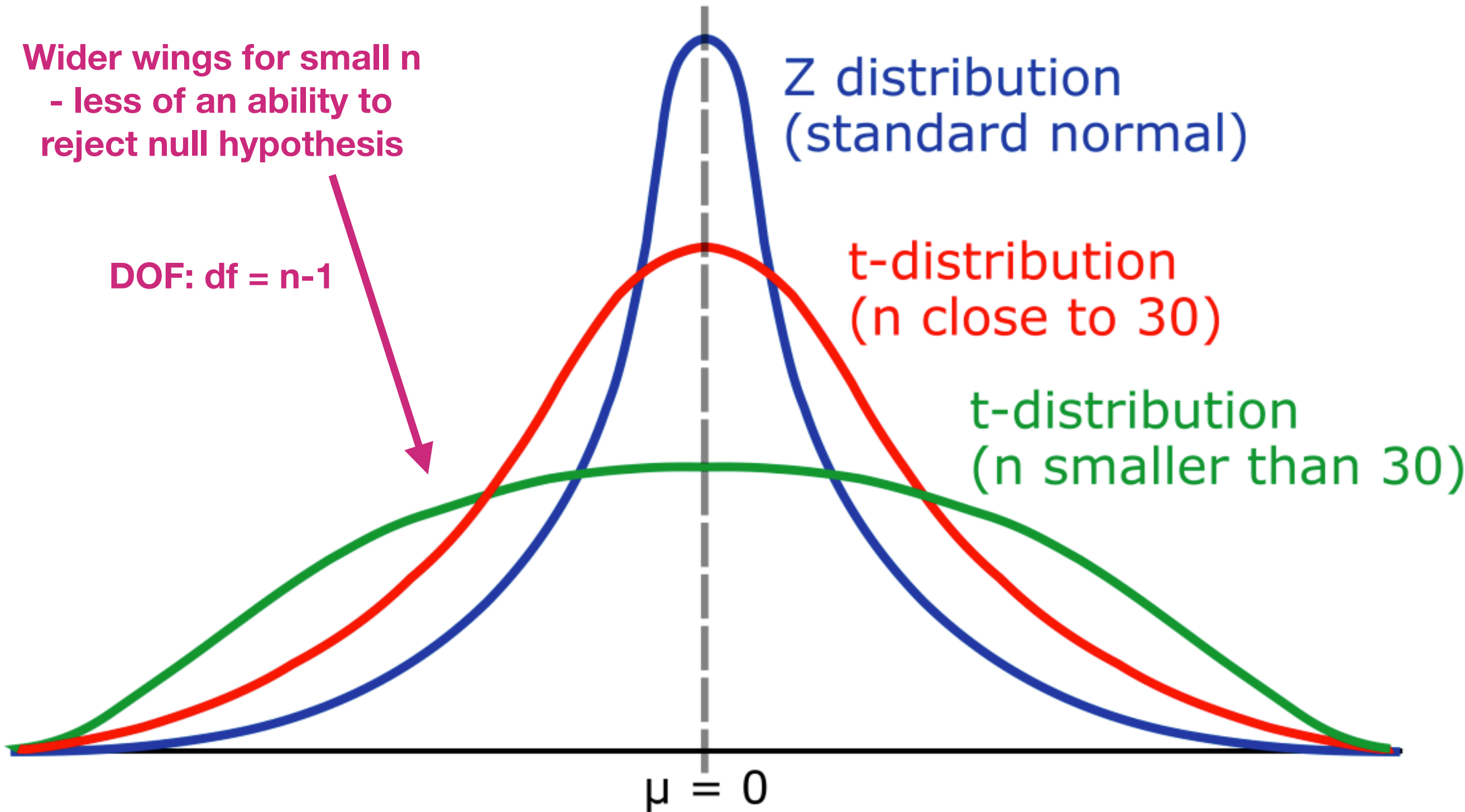
DOF:  $df = n - 1$

Z distribution  
(standard normal)

t-distribution  
(n close to 30)

t-distribution  
(n smaller than 30)

$\mu = 0$



# **An aside on DOF (degrees of freedom)**



# An aside on DOF (degrees of freedom)

Lets say you're into hats and have 1 for every day of the week...



but! you don't want to wear the same hat in the same week (fashion after all)

Day 1: 7 DOF

Day 2: 7-1 DOF

Day 3: 7-2 DOF

...

Day 7: only 1 choice: 7-7 DOF

Now say you're doing a t-test to estimate the mean from a sample of 10 data points

Say mean = 3.5, so sum of all 10 points =  $10 \times 3.5 = 35$

Data point 1: can be anything, chose value D1

Data point 2:  $D2 = 35 - (D1 + \text{sum of all other numbers})$

Data point 3:  $D3 = 35 - (D2 + D1 + \text{sum(others)})$

...

Data point 10:  $D10 = 35 - (D9 + D8 + \dots + D1)$

Since D10 is "fixed" by other points:  $\text{DOF} = 10 - 1 = n - 1$

**DOF = #data points - #constraints**

Wider wings for small n  
- less of an ability to  
reject null hypothesis

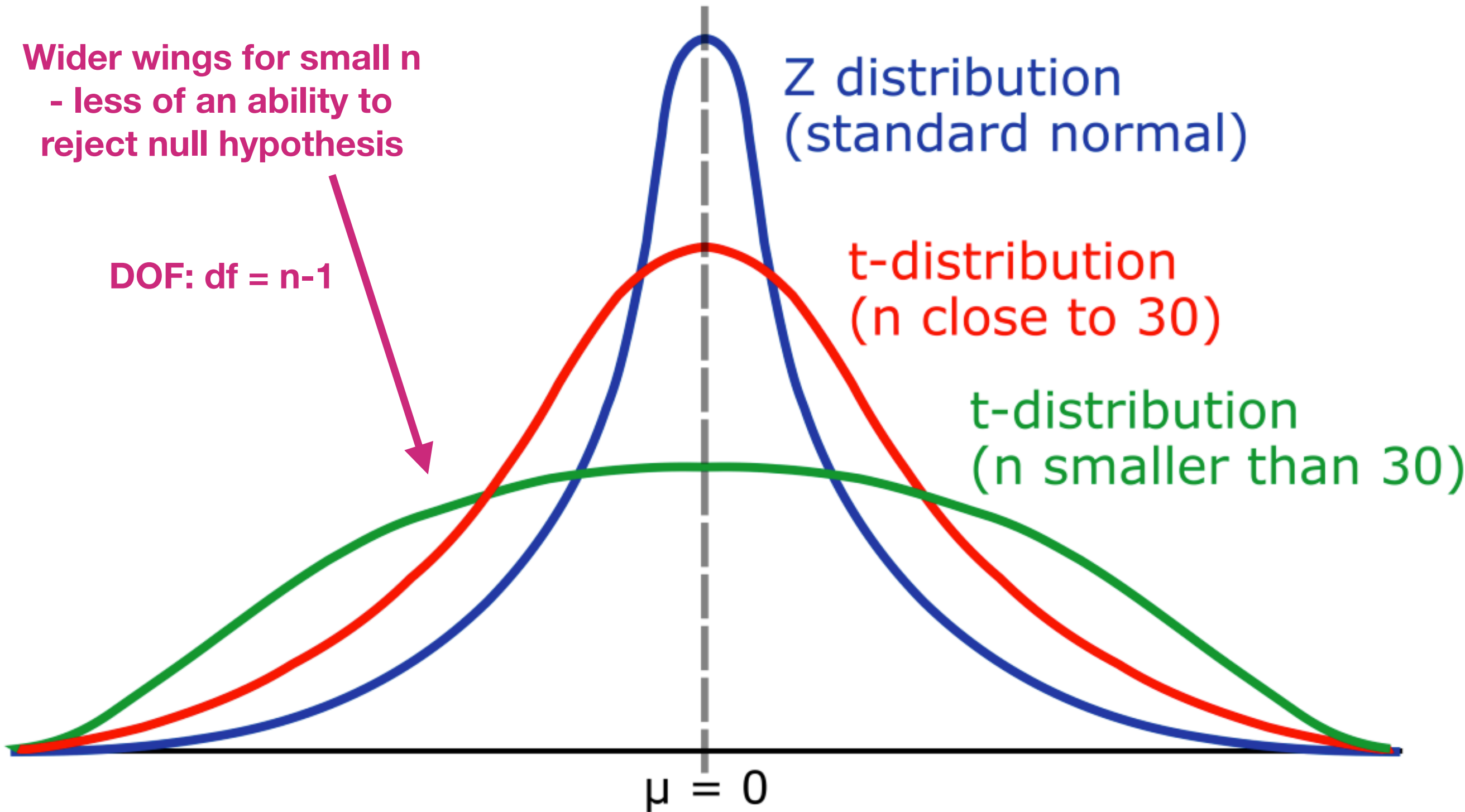
DOF:  $df = n - 1$

Z distribution  
(standard normal)

t-distribution  
(n close to 30)

t-distribution  
(n smaller than 30)

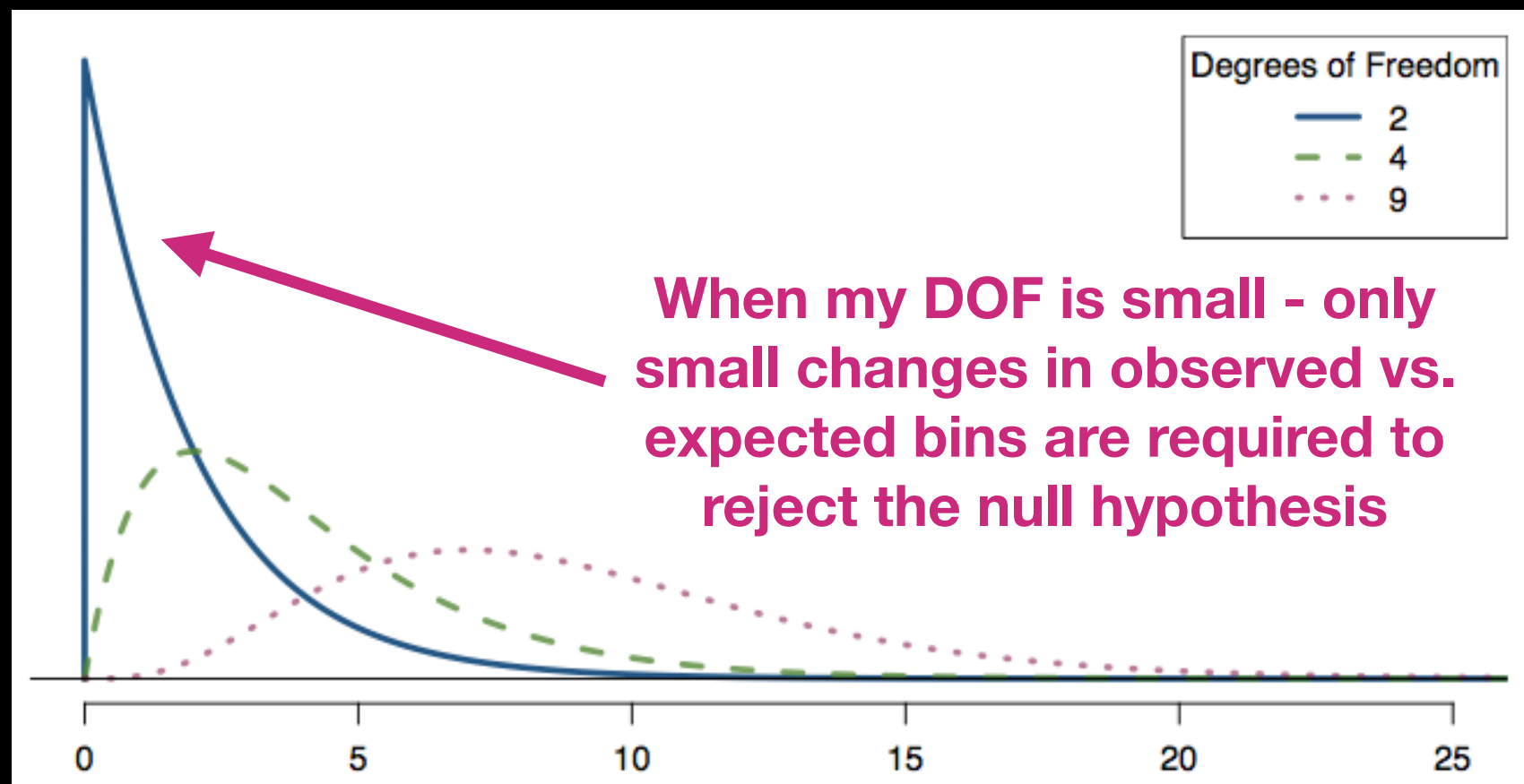
$\mu = 0$



# The chi-square distribution

In order to determine if the  $\chi^2$  statistic we calculated (24.73) 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.



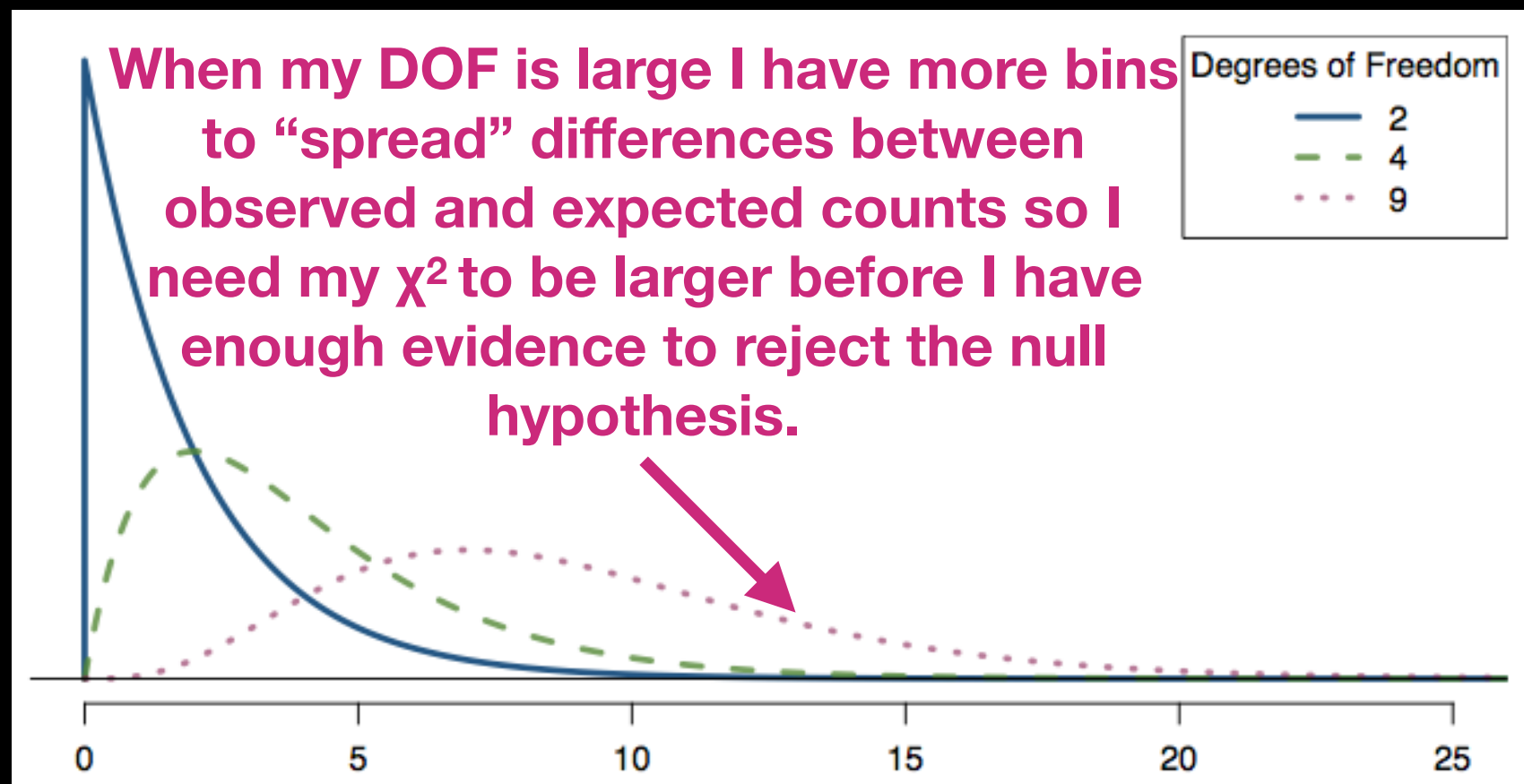
$$\text{df} = k - 1$$

$k = \# \text{ of bins}$

# The chi-square distribution

In order to determine if the  $\chi^2$  statistic we calculated (24.73) 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.



$$df = k - 1$$

**k = # of bins**

# Back to Labby's dice

The research question was: Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

The hypotheses were:

$H_0$ : 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.

We had calculated a test statistic of  $\chi^2 = 24.67$ .

All we need is the df and we can calculate the tail area (the p-value) and make a decision on the hypotheses.

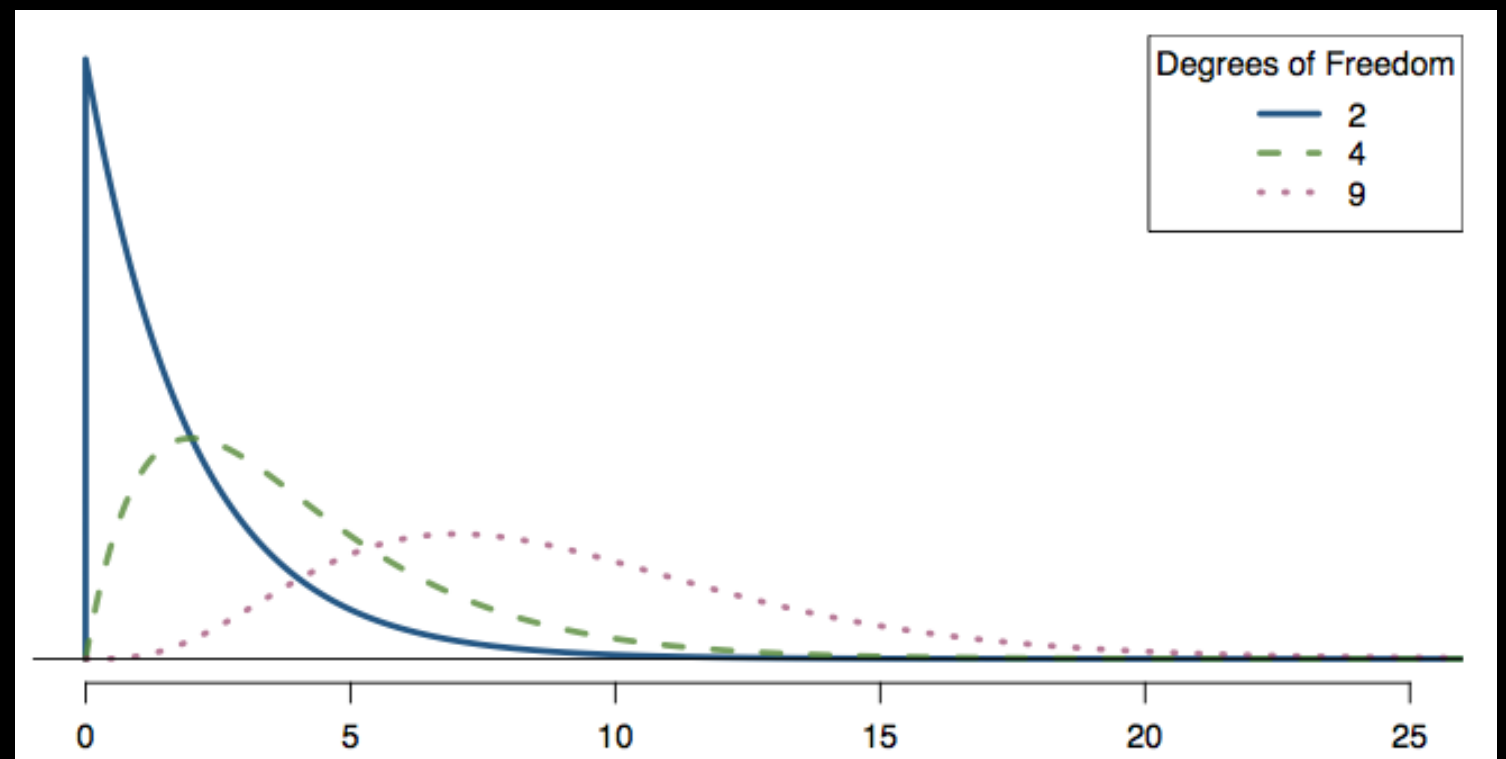
# Degrees of freedom for a goodness of fit test

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

$$df = 6 - 1 = 5$$



# Finding a 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**.



(more on how to do this in R in a few slides)

**p-value < 0.05 (our typical level of significance)**

*Reject  $H_0$ , the data provide convincing evidence that the dice are biased.*

# 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.

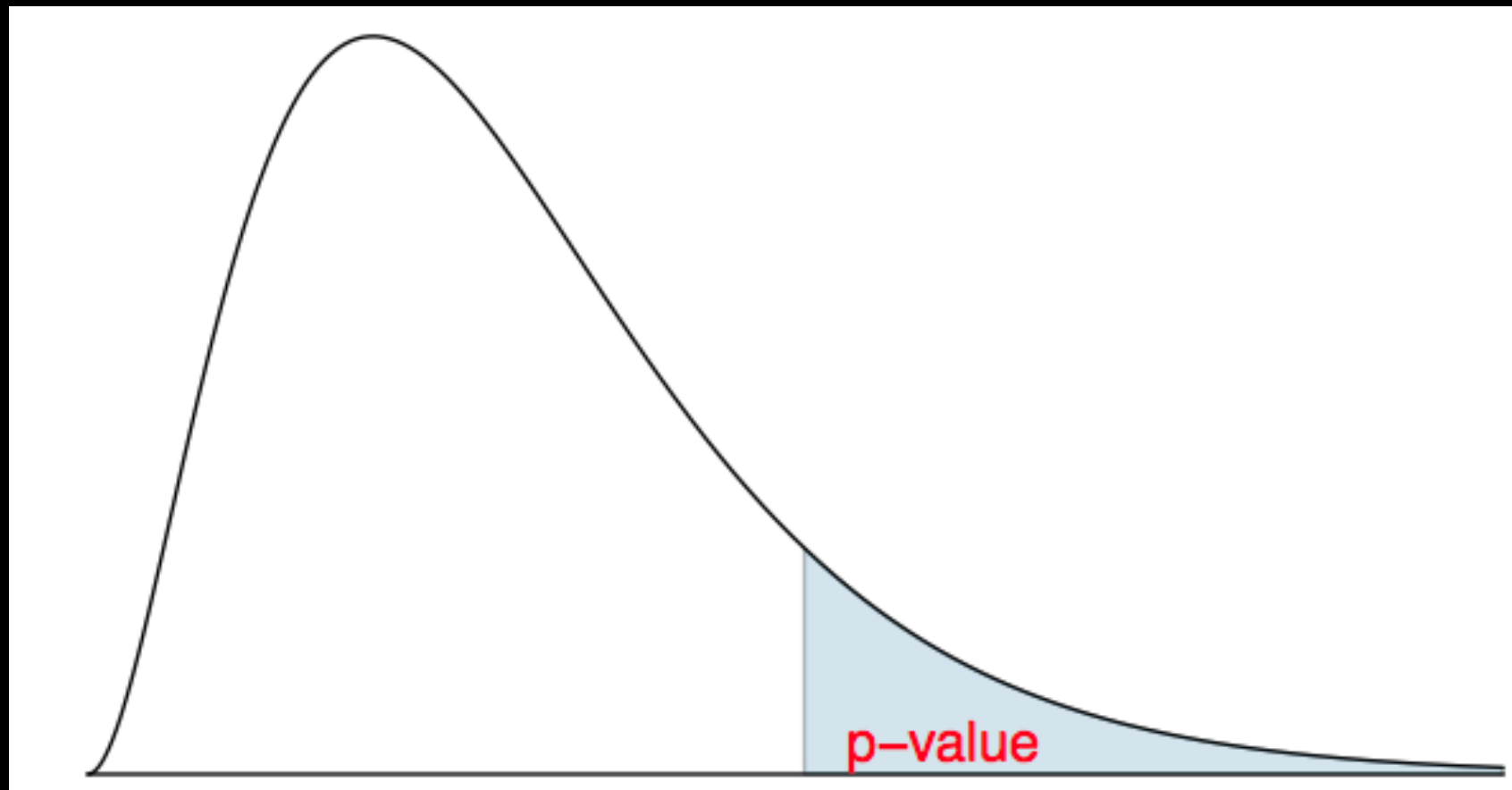




# Recap: p-value for a chi-square test

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.



# Conditions for the chi-square test

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

**Sample size:** Each particular scenario (i.e. cell) must have at least 5 expected cases.

**$df > 1$ :** Degrees of freedom must be greater than 1.

Failing to check conditions may unintentionally affect the test's error rates.

**R-Practice!**

# Comparing means with ANOVA

# z/t test vs. ANOVA - Purpose

## Z or T test (normal or t-dist)

Compare means from two groups to see whether they are so far apart that the observed difference cannot reasonably be attributed to sampling variability.

$$H_0: \mu_1 = \mu_2$$

## ANOVA

Compare the means from two or more groups to see whether they are so far apart that the observed differences cannot all reasonably be attributed to sampling variability.

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

# z/t test vs. ANOVA - Method

## Z or T test

Compute a test statistic  
(a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

## ANOVA

Compute a test statistic  
(a ratio).

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

# z/t test vs. ANOVA - Method

# Z or T test

Compute a test statistic  
(a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

# ANOVA

Compute a test statistic  
(a ratio).      **mean square between groups**  
                                 **variability bet groups**      **(MSG)**

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

## mean square error (MSE)

$$MSG = \frac{1}{df_G}SSG = \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$$

# z/t test vs. ANOVA - Method

## Z or T test

Compute a test statistic  
(a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

Large test statistics lead to small p-values.

If the p-value is small enough  $H_0$  is rejected, we conclude that the population means are not equal.

In order to be able to reject  $H_0$ , we need a small p-value, which requires a large F statistic.

In order to obtain a large F statistic, variability between sample means needs to be greater than variability within sample means.

## ANOVA

Compute a test statistic  
(a ratio).      mean square between groups  
(MSG)

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

mean square error (MSE)



# ANOVA - Conditions

- (1) Independence: all sample sets have randomly sampled data
- (2) Normality: all sample sets are normally distributed
- (3) Constant Variance: all sample sets have the same variance

**Let's do an ANOVA with R.**

# Which means differ?

Last class we discussed the difference of two means - but what if we want to know about several means? The natural question that follows is “which ones?”

We can do two sample t tests for differences in each possible pair of groups.

Can you see any pitfalls with this approach?

- When we run too many tests, the Type 1 Error rate increases.
- This issue is resolved by using a modified significance level.

		Decision	
		fail to reject $H_0$	reject $H_0$
Truth	$H_0$ true	✓	Type 1 Error
	$H_A$ true	Type 2 Error	✓

# Multiple comparisons

The scenario of testing many pairs of groups is called **multiple comparisons**.

The **Bonferroni correction** suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^* = \alpha / K$$

where K is the number of comparisons being considered.

If there are k groups, then usually all possible pairs are compared and  $K = k * (k - 1) / 2$ .

## Let's test this out with R.

**Next week:**

**Linear Regression: Beginning ML**

# Linear Regression: Beginning ML - Where are we going with this?

Basic questions:

What is the underlying model of our data?

How accurate are the predictions we make from this model?

