Chapter 6: Inference for categorical data

Wen-Han Hwang
(Slides primarily developed by Mine Çetinkaya-Rundel from OpenIntro.)

Institute of Statistics
National Tsing Hua University
Taiwan

OpenIntro

Outline

Inference for a single proportion

Sampling distribution of $\hat{p}$

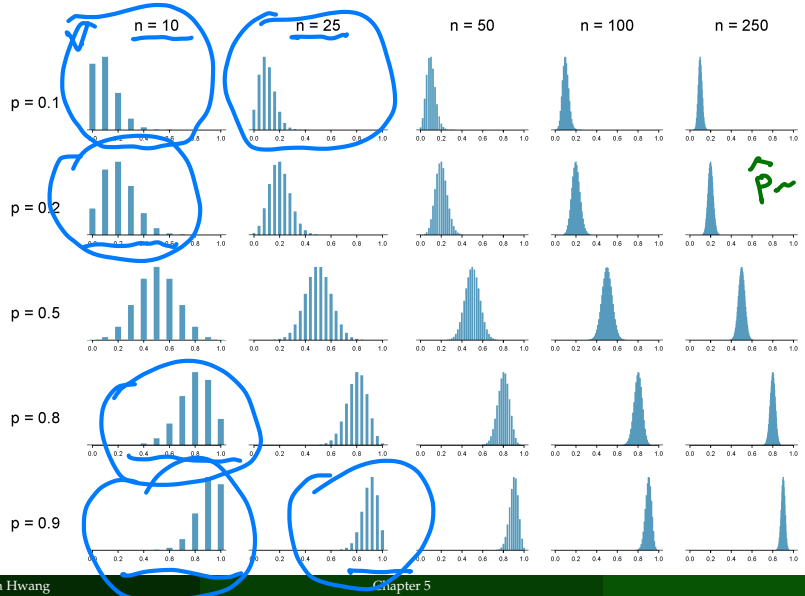
The sampling distribution for $\hat{p}$ based on a sample of size n from a population with a true proportion p is nearly normal when:

1. The sample's observations are independent, e.g. are from a simple random sample.
2. We expected to see at least 5 successes and 5 failures in the sample, i.e. $np \geq 5$ and $n(1-p) \geq 5$. This is called the **success-failure condition.**

When these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error $SE = \sqrt{\frac{p(1-p)}{n}}$.

What happens when np and/or $n(1-p) < 5$?

Dist. of $\hat{p}$. CLT



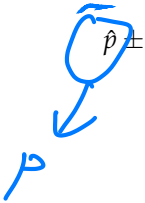
$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

Confidence intervals for a proportion

- Normal Approx. for the dist of $\hat{p}$

A confidence interval provides a range of plausible values for the parameter p, and when $\hat{p}$ can be modeled using a normal distribution, the confidence interval for p takes the form

$$\hat{p} \pm z^\star \times SE$$



Example

A simple random sample of 826 payday loan borrowers was surveyed to better understand their interests around regulation and costs. 70% of the responses supported new regulations on payday lenders.

Construct a 95% confidence interval for p, the proportion of payday borrowers who support increased regulation for payday lenders.

Using the point estimate 0.70, $z^\star = 1.96$ for a 95% confidence interval, and the standard error $SE = 0.016$, the confidence interval is

$$\text{point estimate} \pm z^\star \times SE \quad \rightarrow \quad 0.70 \pm 1.96 \times 0.016 \quad \rightarrow \quad (0.669, 0.731)$$

Hypothesis testing for a proportion

One possible regulation for payday lenders is that they would be required to do a credit check and evaluate debt payments against the borrower's finances. We would like to know: would borrowers support this form of regulation?

Set up hypotheses to evaluate whether borrowers have a majority support or majority opposition for this type of regulation.

$p = 51\%$

Do payday loan borrowers support a regulation that would require lenders to pull their credit report and evaluate their debt payments? From a random sample of 826 borrowers, 51% said they would support such a regulation.

Example

$H_0 : p = 0.5$, $H_1 : p > 0.5$, evaluate whether the poll provides convincing evidence that a majority of payday loan borrowers support a new regulation that would require lenders to pull credit reports and evaluate debt payments.

With hypotheses already set up and conditions checked, we can move onto calculations.The standard error in the context of a proportion hypothesis test is computed using the null value, $p_0 = 0.5$:

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{826}} = 0.017$$

Example (cont.)

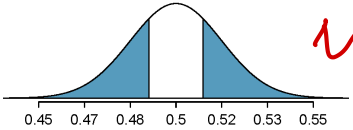


Figure: A normal distribution is shown with a center of 0.5 and a standard deviation of 0.017. Two tails are shaded:

Based on the normal model, the test statistic can be computed as the Z-score of the point estimate:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.51 - 0.50}{0.017} = 0.59$$

The single-tail area is 0.2776, which corresponds to the p-value. Given that the p-value exceeds 0.05, we fail to reject H_0. Consequently, the poll does not offer substantial evidence to conclude that a majority of payday loan borrowers are in favor of implementing regulations for credit checks and debt payment evaluations.

It should be noted that if $H_0 : p = 0.5$ and $H_1 : p \neq 0.50$, the p-value, which accounts for both tail areas, is 0.5552.

Choosing a sample size when estimating a proportion

A university newspaper is conducting a survey to determine what fraction of students support a \$200 per year increase in fees to pay for a new football stadium. How big of a sample is required to ensure the margin of error is smaller than 0.04 using a 95% confidence level?

The margin of error for a sample proportion is

$$z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Our goal is to find the smallest sample size n so that this margin of error is smaller than 0.04. For a 95% confidence level, the value $z^\star$ corresponds to 1.96:

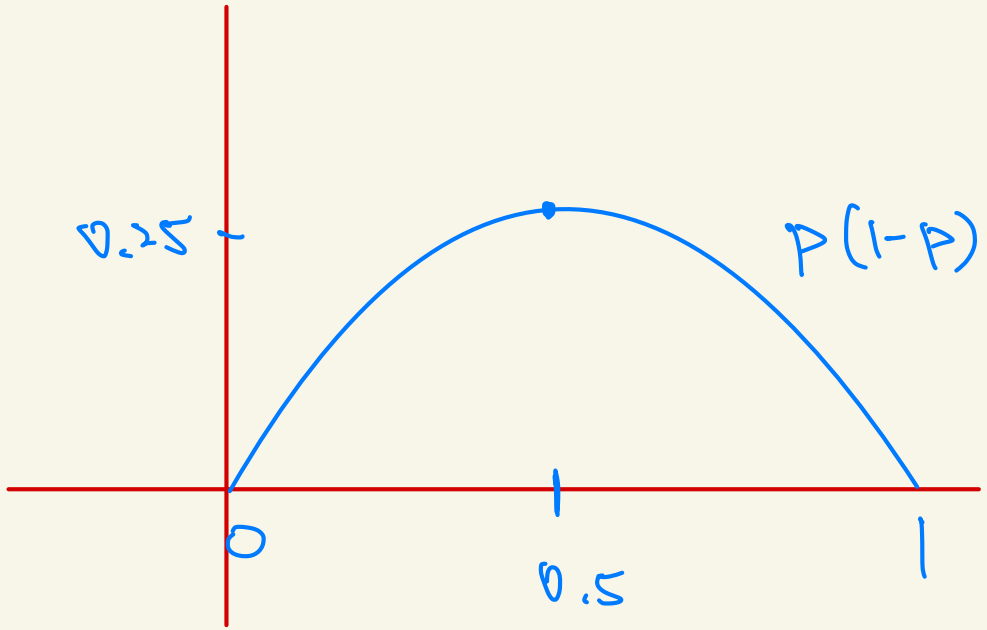
$$1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < 0.04$$

However, the estimate $\hat{p}$ is unknown to us before conducting the survey. We typically adopt the *worst case value* $\hat{p} = 0.5$, since $p(1-p) \leq 0.25$ for all $p \in (0,1)$, which corresponds to the maximum of the function $p(1-p)$.

It follows that the margin of error is maximized when $p = 0.5$. Therefore, we derive:

$$\begin{aligned} 1.96 \times \sqrt{\frac{0.5 \times (1-0.5)}{n}} &< 0.04, \\ 1.96^2 \times \frac{0.25}{n} &< 0.0016, \\ n &> 600.25. \end{aligned}$$

To ensure the sample proportion is within 0.04 of the true proportion with 95% confidence, we require more than 601 participants.



$$1.96^2 \cdot \frac{1}{4n} \le \varepsilon^2$$

$$n \ge \frac{1}{\varepsilon^2}$$

ε	0.04	0.03	0.02	0.01
n	601	1067	? 601.4	600.16

Recap - inference for one proportion

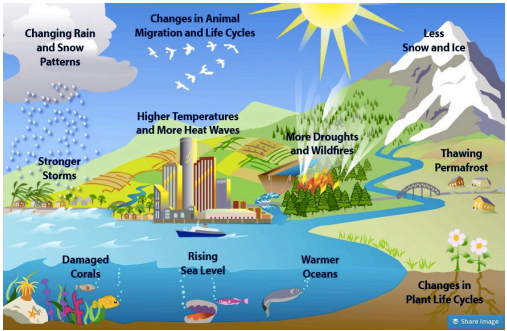
- Population parameter: p, point estimate: $\hat{p}$
- Conditions:
  - independence
    - random sample
  - at least 5 successes and failures
    - if not $\rightarrow$ randomization
- Standard error: $SE = \sqrt{\frac{p(1-p)}{n}}$
  - for CI: use $\hat{p}$
  - for HT: use p_0 $\quad (H_0 : p = p_0)$

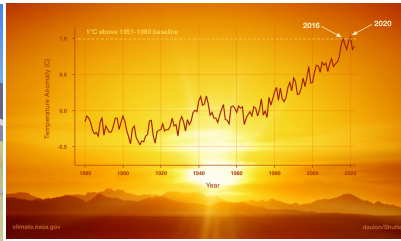
Difference of two proportions

Melting ice cap

Scientists predict that global warming may have big effects on the polar regions within the next 100 years. One of the possible effects is that the northern ice cap may completely melt. Would this bother you a great deal, some, a little, or not at all if it actually happened?

- (a) A great deal
- (b) Some
- (c) A little
- (d) Not at all





Results from the GSS

The General Society Survey (GSS) asks the same question, below are the distributions of responses from the 2010 GSS as well as from a group of introductory statistics students at Duke University:

	GSS	Duke
A great deal	454	69
Some	124	30
A little	52	4
Not at all	50	2
Total	680	105

Parameter and point estimate

- Parameter of interest: Difference between the proportions of all Duke students and all Americans who would be bothered a great deal by the northern ice cap completely melting.

$$p_{Duke} - p_{US}$$

- Point estimate: Difference between the proportions of sampled Duke students and sampled Americans who would be bothered a great deal by the northern ice cap completely melting.

	GSS	Duke
A great deal	454	69
Some	124	30
A little	52	4
Not at all	50	2
Total	680	105

$$\hat{p}_{Duke} - \hat{p}_{US}$$

$\frac{69}{105} - \frac{454}{680}$

$.657 = \hat{p}_1$ $\hat{p}_2 = .668$

Inference for comparing proportions

- The details are the same as before...
- CI: *point estimate* $\pm$ *margin of error* $2 \cdot SE$
- HT: Use $Z = \frac{\textit{point estimate} - \textit{null value}}{SE}$ to find appropriate p-value.
- We just need the appropriate standard error of the point estimate ($SE_{\hat{p}_{Duke} - \hat{p}_{US}}$), which is the only new concept.

Standard Error of the Difference Between Two Sample Proportions

Given the standard error of the difference between two sample proportions can be expressed as:

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

In practice, $\hat{p}_1$ and $\hat{p}_2$ are used as estimators for p_1 and p_2 respectively.

$X \perp Y$

$$Var(aX+bY) = a^2 Var(X) + b^2 Var(Y)$$

$$\begin{cases} X \sim B(n_1, P_1) & Var\left(\frac{X}{n_1}\right) = \frac{P_1(1-P_1)}{n_1} \\ Y \sim B(n_2, P_2) & Var\left(\frac{Y}{n_2}\right) = \frac{P_2(1-P_2)}{n_2} \\ Var\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) = Var\left(\frac{X}{n_1}\right) + Var\left(\frac{Y}{n_2}\right) \end{cases}$$

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap ($p_{Duke} - p_{US}$).

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z^\star \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}}$$

$$= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}}$$

$$- 0.011 \pm 0.097$$

$$= (-0.108, 0.086)$$

← Cannot reject $H_0: P_1 = P_2$

Which of the following is the correct set of hypotheses for testing if the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do?

(a) $H_0 : p_{Duke} = p_{US}$
$H_A : p_{Duke} \neq p_{US}$

(b) $H_0 : \hat{p}_{Duke} = \hat{p}_{US}$
$H_A : \hat{p}_{Duke} \neq \hat{p}_{US}$

(c) $H_0 : p_{Duke} - p_{US} = 0$
$H_A : p_{Duke} - p_{US} \neq 0$

(d) $H_0 : p_{Duke} = p_{US}$
$H_A : p_{Duke} < p_{US}$

(a) ≡ (c) equivalent

Both (a) and (c) are correct.

Pooled estimate of a proportion

- In the case of comparing two proportions where $H_0 : p_1 = p_2$, there isn't a given null value we can use to calculated the expected number of successes and failures in each sample.
- Therefore, we need to first find a common (pooled) proportion for the two groups, and use that in our analysis.
- This simply means finding the proportion of total successes among the total number of observations.

$$SE^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Pooled estimate of a proportion

$$H_0 : SE^2 = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2} \quad (p_1 = p_2 = p)$$

$$\hat{p}_c = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$

Calculate the estimated pooled proportion of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion ($\hat{p}_{Duke}$ or $\hat{p}_{US}$) the pooled estimate is closer to? Why?

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$$\hat{p} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$
$$= \frac{69 + 454}{105 + 680} = \frac{523}{785} = 0.666$$

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

Data	Duke	US
A great deal	69	454
Not a great deal	36	226
Total	105	680
$\hat{p}$	0.657	0.668

$(p_1 - p_2 \text{ @ } H_0)$

$$Z = \frac{(\hat{p}_{Duke} - \hat{p}_{US}) - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}} \rightarrow \hat{p}_c$$

$$= \frac{(0.657 - 0.668) - 0}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22$$

$$p - value = 2 \times P(Z < -0.22) = 2 \times 0.41 = 0.82$$

Recap - comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- Conditions:
  - independence within groups
    - random sample
  - independence between groups
  - at least 5 successes and failures in each group
    - if not $\rightarrow$ randomization (Section 6.4)
- $SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - for CI: use $\hat{p}_1$ and $\hat{p}_2$
  - for HT:
    - when $H_0 : p_1 = p_2$: use $\hat{p}_{pool} = \frac{\#\, suc_1 + \#suc_2}{n_1 + n_2}$
    - when $H_0 : p_1 - p_2 =$ *(some value other than 0)*: use $\hat{p}_1$ and $\hat{p}_2$
      - this is pretty rare

Reference - standard error calculations

	one sample	two samples
proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

- When working with proportions,
  - if doing a hypothesis test, p comes from the null hypothesis
  - if constructing a confidence interval, use $\hat{p}$ instead

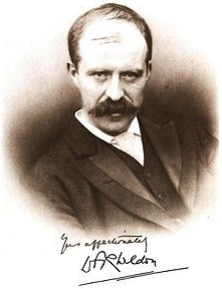
one sample $p = p_0$,

two samples $p_1 = p_2 = p_c$

Chi-square test of GOF

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.

Table 2—Current Data on Dice: 26,306 Throw

Number of Successes	Observed Frequency	Theoretical Frequency $p = 1/3$	
0	216	203	
1	1194	1216	
2	3292	3345	
3	5624	5576	
4	6186	6273	
5	5047	5018	
6	2953	2927	
7	1288	1255	
8	406	392	399
9	85	87	89
10	13	13	13
11	2	1	1
12	0	0	0
Total	**26,306**	$\chi^2_{[10]} = 5.62$	$\chi^2_{[9]} = 4.32$

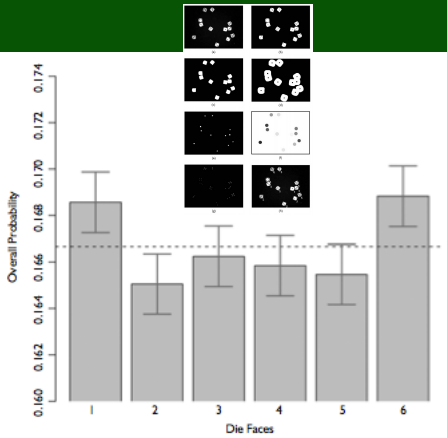
Note: A die was considered a success if five or six pips were showing.

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=95EErdou02w`

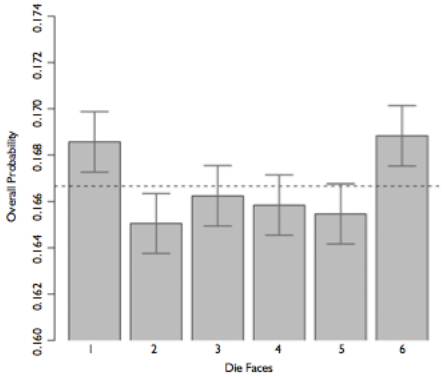
- The rolling-imaging process took about 20 seconds per roll.



- 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) $\frac{12 \times 26{,}306}{6} = 52,612$

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

Why are the expected counts the same for all outcomes but the observed counts are different? 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.The observed counts follow the same distribution as the expected counts ($p_1 = \cdots = p_6$).

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 (p_i are not equal).

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.

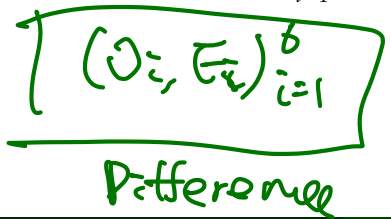
  These two ideas will help in the construction of an appropriate test statistic for count data.

Chi-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 (χ^2) statistic.

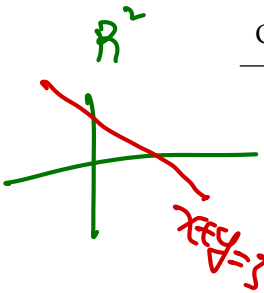
χ^2 statistic

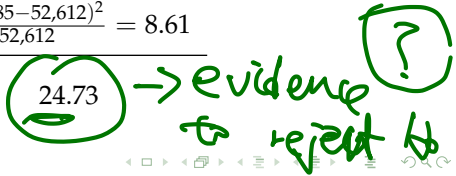
$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$ where k = total number of cells



Calculating the chi-square statistic

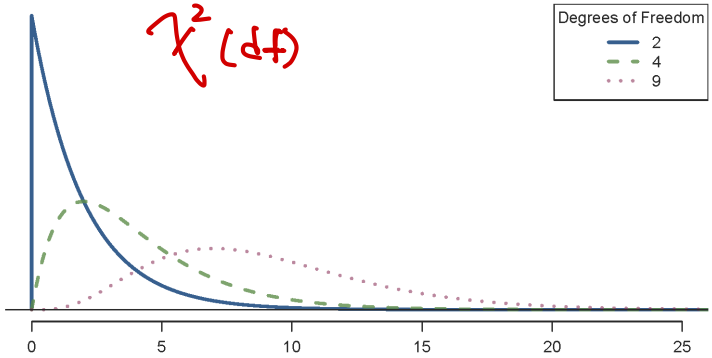
Outcome	Observed	Expected	$\frac{(O-E)^2}{E}$
1	53,222	52,612	$\frac{(53,222-52,612)^2}{52,612} = 7.07$
2	52,118	52,612	$\frac{(52,118-52,612)^2}{52,612} = 4.64$
3	52,465	52,612	$\frac{(52,465-52,612)^2}{52,612} = 0.41$
4	52,338	52,612	$\frac{(52,338-52,612)^2}{52,612} = 1.43$
5	52,244	52,612	$\frac{(52,244-52,612)^2}{52,612}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





The chi-square distribution

- 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.



As the df increases,

(a) the center of the χ^2 distribution increases as well

(b) the variability of the χ^2 distribution increases as well

(c) the shape of the χ^2 distribution becomes more like a normal

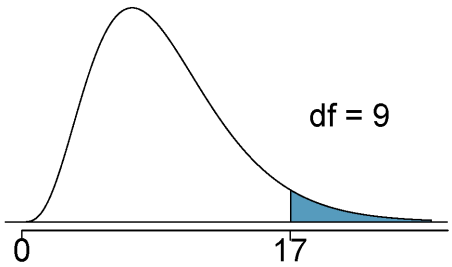
all correct

Finding areas under the chi-square curve

- p-value = tail area under the chi-square distribution (as usual)
- For this we can use technology, or a chi-square probability table.

Finding areas under the chi-square curve (cont.)

Estimate the shaded area (above the cutoff value of 17) under the χ^2 curve with $df = 9$.

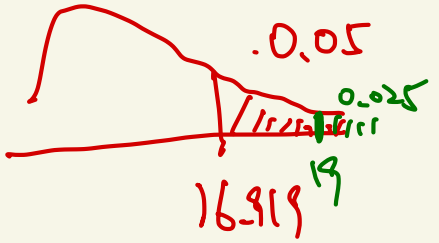


(a) 0.05
(b) 0.02
(c) between 0.02 and 0.05
(d) between 0.05 and 0.1
(e) between 0.01 and 0.02

```
> pchisq( 17, df = 9, lower.tail = FALSE)
[1] 0.04871598
```

( ) = 0.952

Chi-Square Right-Tail Probability ($\geq \chi^2$)										
DF	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169



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.

  $H_0: P_1 = \cdots = P_6$

  H_1: P_i are not equal
- 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.

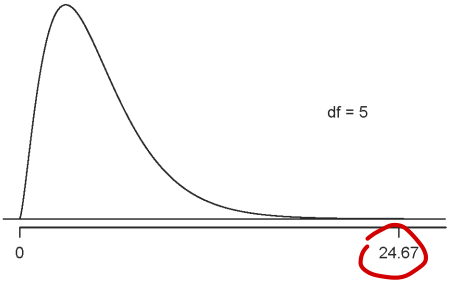
$$df = k - 1$$

- 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.



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

Chi-Square Right-Tail Probability ($\geq \chi^2$)										
DF	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801

Conclusion of the hypothesis test

We calculated a p-value less than 0.001. 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 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.
- Casino dice are designed with flush faces, where the pips are filled with a material of identical density to the rest of the dice surface, ensuring precise balance and fairness in games.

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	p_1	63.29%	p_{10}
(2) Mousavi	136	p_2	34.10%	p_{20}
(3) Minor candidates	30	p_3	2.61%	p_{30}
Total	504		100%	
	↓ observed		↓ expected distribution	

H_0: $p_1 = p_{10}$, $p_2 = p_{20}$, $p_3 = p_{30}$

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

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

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

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

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

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

p-value $<<< 0.005$

p-value < 0.0005

Chi-Square Right-Tail Probability ($\geq \chi^2$)										
DF	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801

Conclusion

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.

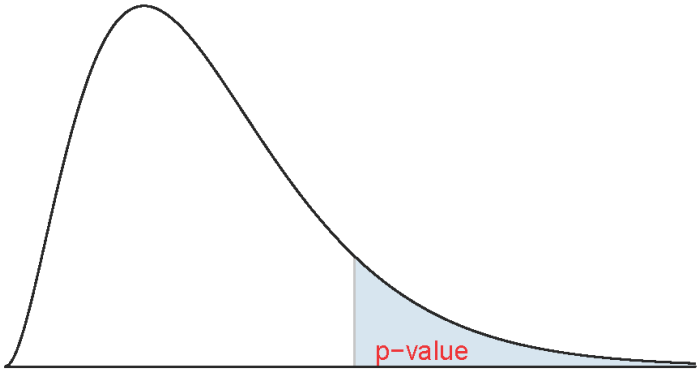
(b) p-value is high, H_0 is not rejected. The observed counts from the poll follow the same distribution as the reported votes.

(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.

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.



Chi-square test of independence

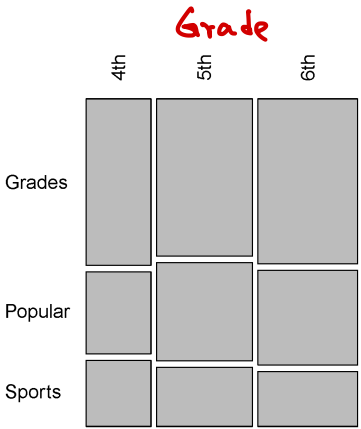
Popular kids

In the dataset `popular`, students in grades 4-6 were asked whether good grades, athletic ability, or popularity was most important to them. A two-way table separating the students by grade and by choice of most important factor is shown below. Do these data provide evidence to suggest that goals vary by grade?

	Grades	Popular	Sports
4^{th}	63	31	25
5^{th}	88	55	33
6^{th}	96	55	32

H_0: Goal ⊥ Grade

H_1



Chi-square test of independence

- The hypotheses are:

  H_0: Grade and goals are independent. Goals do not vary by grade.
  H_A: Grade and goals are dependent. Goals vary by grade.

- The test statistic is calculated as

$$\chi^2_{df} = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \quad \text{where} \quad df = (R-1) \times (C-1),$$

  where k is the number of cells, R is the number of rows, and C is the number of columns.
  Note: We calculate df differently for one-way and two-way tables.

- The p-value is the area under the χ^2_{df} curve, above the calculated test statistic.

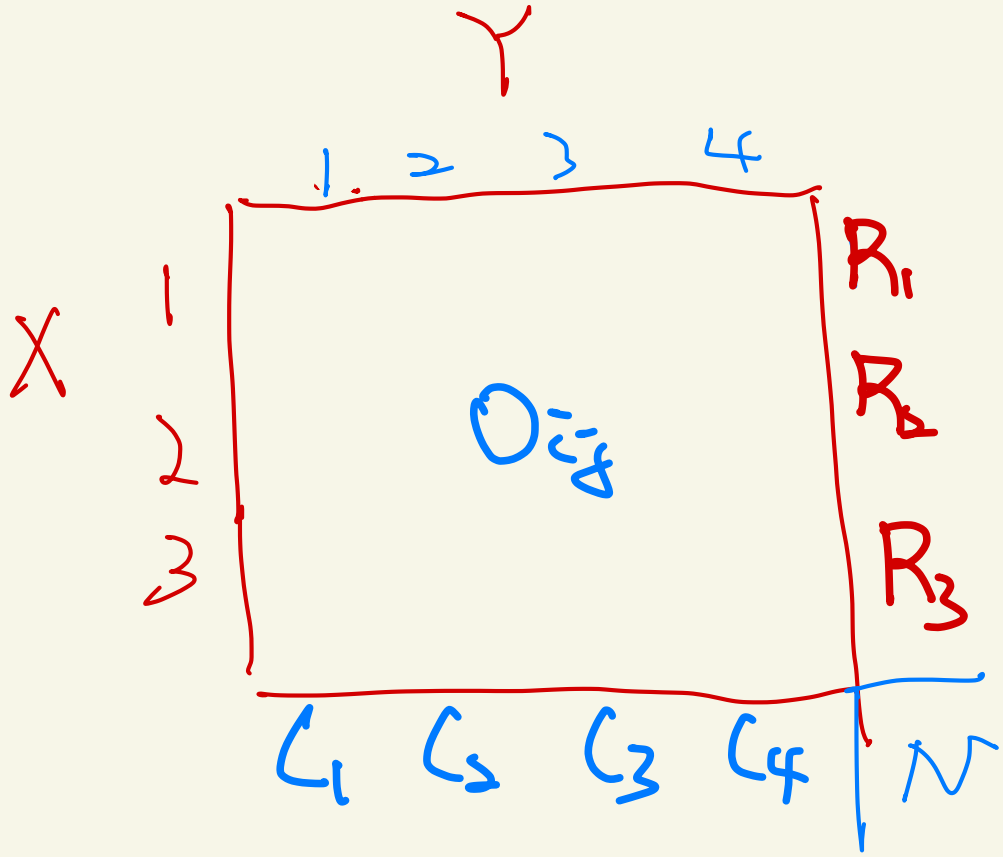
Expected counts in two-way tables

Expected counts in two-way tables

$$\text{Expected Count} = \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$$

	Grades	Popular	Sports	Total
4^{th}	63	31	25	119
5^{th}	88	55	33	176
6^{th}	96	55	32	183
Total	247	141	90	478

$$E_{row\ 1,col\ 1} = \frac{119 \times 247}{478} = 61 \qquad E_{row\ 1,col\ 2} = \frac{119 \times 141}{478} = 35$$



$X \perp Y$

$\Leftrightarrow$

$P(X=i, Y=j) = P(X=i)P(Y=j)$

$$E_{ij} = N \cdot P(X=i, Y=j) = N P(X=i) P(Y=j)$$

$$\hat{P}(X=i) = \frac{R_i}{N}$$

$$\hat{P}(Y=j) = \frac{C_j}{N}$$

$$\therefore E_{ij} = N \cdot \frac{R_i}{N} \cdot \frac{C_j}{N} = \frac{R_i \cdot C_j}{N}$$

$\sum_{i=1}^{r} R_i/N = 1$, $df = (r-1)$, $\sum_{j=1}^{c} \frac{C_j}{N} = 1$, $df = (c-1)$

If $P(X=i, Y=j) = P(X=i)\,P(Y=j)$

$P(X=i \mid X=1) = P(X=i)$ $\forall i$

$P(Y=j \mid X=2) = P(Y=j)$ $\forall j$

Calculating the test statistic in two-way tables

Expected counts are shown in blue next to the observed counts.

	Grades	Popular	Sports	Total
4^{th}	63 61	31 35	25 23	119
5^{th}	88 91	55 52	33 33	176
6^{th}	96 95	55 54	32 34	183
Total	247	141	90	478

$$\chi^2 = \sum \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \cdots + \frac{(32-34)^2}{34} = 1.3121$$

$$df = (R-1) \times (C-1) = (3-1) \times (3-1) = 2 \times 2 = 4$$

Calculating the p-value

Which of the following is the correct p-value for this hypothesis test?

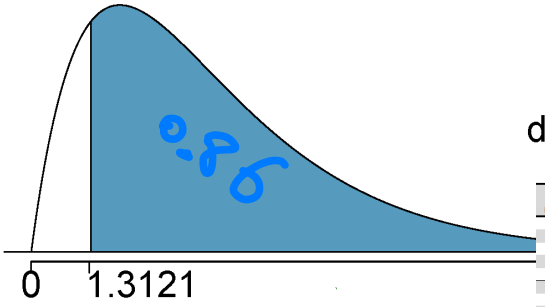
$\chi^2 = 1.3121 \qquad df = 4$

- Enter value for degrees of freedom.
- Enter a value for one, and only one, of the other textboxes.
- Click **Calculate** to compute a value for the remaining textbox.

Degrees of freedom	4
Chi-square value (x)	1.3
Probability: $P(X^2 \le 1.3)$	0.13862
Probability: $P(X^2 \ge 1.3)$	0.86138

Calculate

(a) more than 0.3
(b) between 0.3 and 0.2
(c) between 0.2 and 0.1
(d) between 0.1 and 0.05
(e) less than 0.001



df = 4

P-value > 0.1

Chi-Square Right-Tail Probability ($\ge \chi^2$)

DF	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300

Conclusion

Do these data provide evidence to suggest that goals vary by grade?

H_0: Grade and goals are independent. Goals do not vary by grade.
H_A: Grade and goals are dependent. Goals vary by grade.

Since p-value is high, we fail to reject H_0. The data do not provide convincing evidence that grade and goals are dependent. It doesn't appear that goals vary by grade.

Final Remarks: Proportions and Means

Table: Statistical inferences for Proportions and Means

Measure/Context	Proportions	Means
CI (single sample)	$\hat{p} \pm z \times SE$	$\hat{\mu} \pm z \times SE$
CI (two samples)	$\hat{p}_1 - \hat{p}_2 \pm z \times SE$	$\hat{\mu}_1 - \hat{\mu}_2 \pm z \times SE$
Test Statistic (one-sample test)	$\frac{\hat{p}-p_0}{SE}$	$\frac{\hat{\mu}-\mu_0}{SE}$
Test Statistic (two-sample test)	$\frac{\hat{p}_1-\hat{p}_2}{SE}$	$\frac{\hat{\mu}_1-\hat{\mu}_2}{SE}$
Testing Problem	Single-sample: testing the proportion against a benchmark p_0. Two-sample: comparing two proportions.	Single-sample: testing the mean against a benchmark μ_0. Two-sample: comparing two means.

Final Remarks: Proportions and Means

	Single Sample	Two Samples
Proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
Mean	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

- For proportions:
  - Confidence Interval (CI): $\hat{p} \pm z \times SE$, $\hat{p}_1 - \hat{p}_2 \pm z \times SE$; incorporate $\hat{p}$, $\hat{p}_1$, and $\hat{p}_2$ into the SE computation.
  - Test statistic: $\frac{\hat{p}-p_0}{SE}$, $\frac{\hat{p}_1-\hat{p}_2}{SE}$; use p_0 and a pooled estimate for p in the one-sample and two-sample scenarios, respectively.
- For means:
  - CI: $\hat{\mu} \pm z \times SE$, $\hat{\mu}_1 - \hat{\mu}_2 \pm z \times SE$; where $\hat{\mu} = \bar{x}$, $\hat{\mu}_1 = \bar{x}_1$, and $\hat{\mu}_2 = \bar{x}_2$ are the sample means.
  - Test statistic: $\frac{\hat{\mu}-\mu_0}{SE}$, $\frac{\hat{\mu}_1-\hat{\mu}_2}{SE}$; utilize a pooled estimate for SE in the two-sample case.

- χ^2 dist.

- $Z \sim N(0,1) \Rightarrow Z^2 \sim \chi^2(1)$

- $Z_1 \cdots Z_r \overset{iid}{\sim} N(0,1) \Rightarrow Z_1^2 + \cdots + Z_r^2 \sim \chi^2(r)$

-

$X_1 \dots X_n \overset{i.i.d}{\sim} N(\mu, \sigma)$

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \Rightarrow \frac{n\cdot(\bar{X}-\mu)^2}{\sigma^2} \sim \chi^2(1)$

$\frac{X_1-\mu}{\sigma}, \dots \frac{X_n-\mu}{\sigma} \overset{i.i.d}{\sim} N(0,1)$

$\Rightarrow \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$

Well Known Equality

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{n}{\sigma^2}(\bar{x}-\mu)^2$$

$\downarrow$ $\chi^2(n)$

$x_1 \cdots x_n$ iid $N(\mu, \sigma^2)$

$\downarrow$ $\chi^2(1)$

$\downarrow$ can be shown that

$\chi^2(n-1)$

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$$

$$\sqrt{\frac{\sum (X_i-\bar{X})^2}{\sigma^2}} \sim \chi^2(n-1)$$

$$\sim t(n-1)$$