Last time

- Compare proportion $\hat{p}$ to null value p_0
  - Statistic: Z-score: $Z = \frac{\hat{p} - p_0}{SE}$
  - Under the null, $Z \sim N(0,1)$
- Compare two proportions $\widehat{p_1}, \widehat{p_2}$
- Null value $p_1 - p_2 = 0$
- Statistic: Z-score: $Z = \frac{\widehat{p_1} - \widehat{p_2} \; - p_1 + p_2}{SE_{p_1 - p_2}}$

Practice

A drone company is considering a new manufacturer for rotor blades. The new manufacturer would be more expensive, but they claim their higher-quality blades are more reliable, with more than 3% more blades passing inspection than their competitor. Set up appropriate hypotheses for the test.

Practice

- Identify the research question

- Identify a quantity related to the research question whose value we don't know ('**parameter**').
- Writing the statistical hypotheses in terms of that parameter of interest.

- Collect data and calculate a statistic

- Find the distribution of the statistic under the null hypothesis

- Find the p-value (probability that the result we got or a more extreme one happens just by chance given that the null hypothesis is true).
- Decide if the p-value is small or large
- Reject if p-value is lower than the significance threshold a.

Practice

- Identify the research question (C2 blades are better than C1 blades)
- Identify a quantity related to the research question whose value we don't know ('**parameter**'). $p_2 - p_1$
- Writing the statistical hypotheses in terms of that parameter of interest. H_0: $p_2 - p_1$=0.03 and H_a: $p_2 - p_1 > 0.03$.
- Collect data and calculate a statistic (Z-score:$\frac{\widehat{p_2}-\widehat{p_1}-(\mathrm{p}_2-\mathrm{p}_1)}{SE_{p_2-p_1}} = \frac{\widehat{p_2}-\widehat{p_1}-0.03}{SE_{p_2-p_1}}$)
- Find the distribution of the statistic under the null hypothesis $N(0,1)$
- Find the p-value (probability that the result we got or a more extreme one happens just by chance given that the null hypothesis is true).
- Decide if the p-value is small or large
- Reject if p-value is lower than the significance threshold a.

Practice

The quality control engineer collects a sample of blades, examining 1000 blades from each company, and she finds that 899 blades pass inspection from the current (C1) supplier and 958 pass inspection from the prospective (C2) supplier.

Find the p-value

Should we change suppliers?

Chi-Square test of GOF

Fisher’s exact test

- Ronald Fisher offered lady Muriel Bristol, a cup of tea.
- She declined after watching Fisher prepare it, saying that she preferred the taste when the milk was poured in the cup first.
- Fisher and others scoffed at this and a colleague, William Roach, suggested a test.

Fisher's exact test

- Ronald Fisher offered lady Muriel Bristol, a cup of tea.
- She declined after watching Fisher prepare it, saying that she preferred the taste when the milk was poured in the cup first.
- Fisher and others scoffed at this and a colleague, William Roach, suggested a test.

- 4 cups with milk poured first, 4 cups with milk poured after.
- Otherwise the cups were the same (temperature, appearance etc).

Fisher's exact test

- The lady is offered the tea, and for every cup she guesses:
    - Milk first (MF) or Tea first (TF)

		Guess		
		MF	TF	Total
Prep	MF	4	0	4
	TF	0	4	4
	Total	4	4	8

Contingency table

Fisher's exact test

- The lady is offered the tea, and for every cup she guesses:
  - Milk first (MF) or Tea first (TF)

Once you fix one of the values, all the rest are fixed because the marginals are fixed

		Guess: MF	Guess: TF	Total
Prep	MF	4	0	4
	TF	0	4	4
	Total	4	4	8

Contingency table

Fisher's exact test

- The lady is offered the tea, and for every cup she guesses:
  - Milk first (MF) or Tea first (TF)

- H_0: The lady has no ability of distinguishing the method of preparation (the woman selects randomly).
- x: The number of MF she got right.
- P-value: The probability of observing data at least as extreme (unfavorable to H_0) under the null hypothesis.

		Guess: MF	Guess: TF	Total
Prep	MF	4	0	4
	TF	0	4	4
	Total	4	4	8

Contingency table

Fisher's exact test

- The lady is offered the tea, and for every cup she guesses:
  - Milk first (MF) or Tea first (TF)

- H_0: The lady has no ability of distinguishing the method of preparation (the woman selects randomly).
- x: The number of MF she got right.
- P-value: The probability of observing data at least as extreme (unfavorable to H_0) under the null hypothesis.

- $P\ (X \geq\ x|H_0)$

		Guess MF	Guess TF	Total
Prep	MF	4	0	4
	TF	0	4	4
	Total	4	4	8

Contingency table

$P(X = 4|H_0)$

Fisher's exact test

- Under the null hypothesis, the lady picks 4 cups at random, without replacement, from a population of 4 MF and TF cups
- X: number of MF cups
- $X \sim Hypergeometric(N, K, n)$
  - N is the population size
  - K is the number of success states in the population
  - n is the number of draws
- P(X=x) = $\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$

		Guess		
		MF	TF	Total
Prep	MF	4	0	4
	TF	0	4	4
	Total	4	4	8

Contingency table

$P(X = 4|H_0)$

Fisher's exact test

- Under the null hypothesis, the lady picks 4 cups at random, without replacement, from a population of 4 MF and TF cups
- X: number of MF cups
- $X \sim Hypergeometric(N, K, n)$
  - N is the population size
  - K is the number of success states in the population
  - n is the number of draws
- P(X=x) $= \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$

		Guess		
		MF	TF	Total
Prep	MF	4	0	4
	TF	0	4	4
	Total	4	4	8

Contingency table

$$P(X = 4|H_0) = \frac{1}{70} = 0.014$$

Fisher's exact test

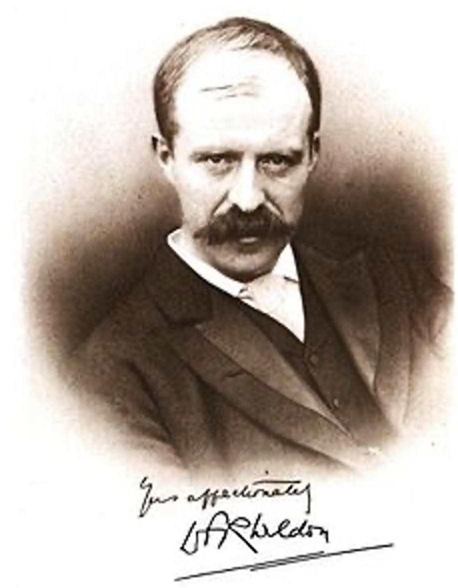
- Under the null hypothesis, the lady picks 4 cups at random, without replacement, from a population of 4 MF and TF cups
- X: number of MF cups
- $X \sim Hypergeometric(N, K, n)$
    - N is the population size
    - K is the number of success states in the population
    - n is the number of draws
- P(X=x) $= \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$

		Guess		
		MF	TF	Total
Prep	MF	3	1	4
	TF	1	3	4
	Total	4	4	8

Contingency table

$$P(X=3|H_0) + P(X=4|H_0) = \frac{16}{70} + \frac{1}{70} = 0.242$$

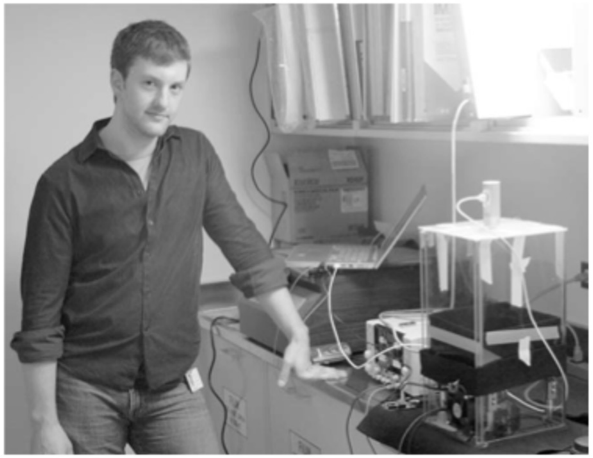
Weldon's dice



- 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

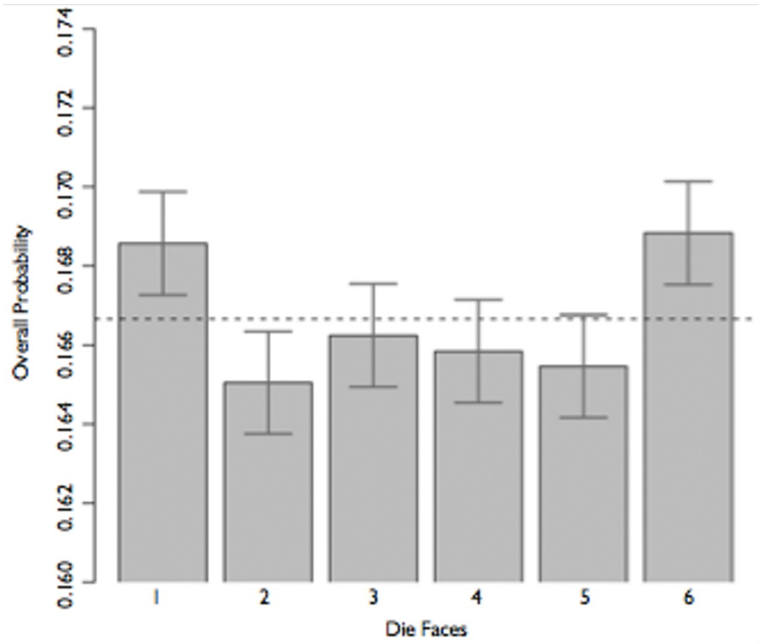


- In 2009, Zacariah Labby (U of Chicago), repeated Weldon's experiment using a homemade dice-throwing, pip counting machine.
  www.youtube.com/watch?v=95EErdouO2w
- The rolling-imaging process took about 20 seconds per roll.

- 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.

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, ..., 6s would he expect to have observed?

(a) 1/6
(b) 12/6
(c) 26,306 / 6
(d) 12 x 26,306 / 6

Expected counts

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

(a) 1/6
(b) 12/6
(c) 26,306 / 6
(d) 12 x 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.

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-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

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

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

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.

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.

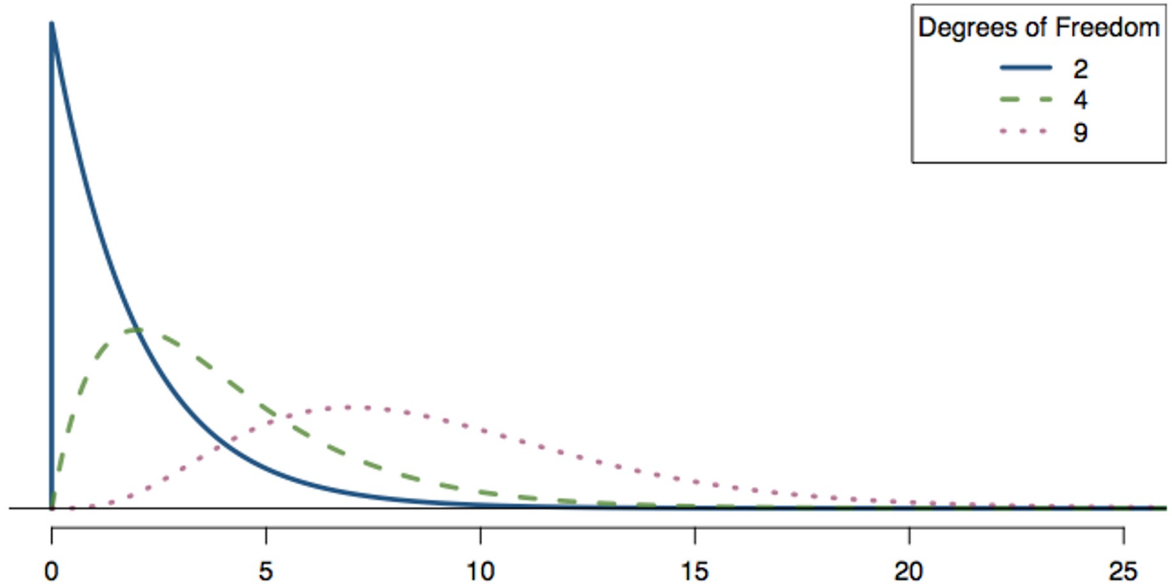
$$\mathrm{X}^2 = \sum_{i=1}^{k} \frac{(N_i - np_i^o)^2}{np_i^0}$$

Under the null, when $n \to \infty, \mathrm{X}^2 \sim \chi^2$ *with k-1 degrees of freedom.*

- The chi-square distribution has just one parameter called *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.

χ^2 distributions

Which of the following is false?



Finding areas under the chi-square curve

- p-value = tail area under the chi-square distribution (as usual)

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

Estimate the shaded area under the chi-square curve with $df = 6$.

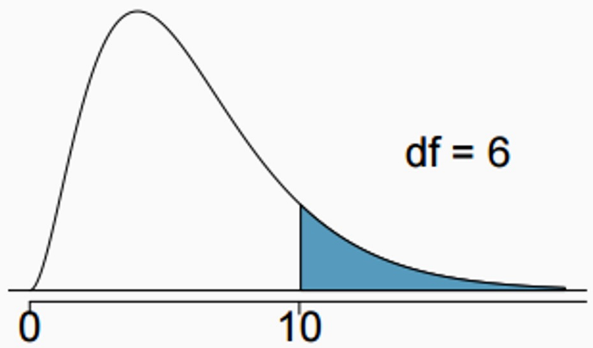
Finding areas under the chi-square curve

Estimate the shaded area under the chi-square curve with *df = 6*.

```
> pchisq(q = 10, df = 6, lower.tail = FALSE)
[1] 0.124652
```

Finding areas under the chi-square curve

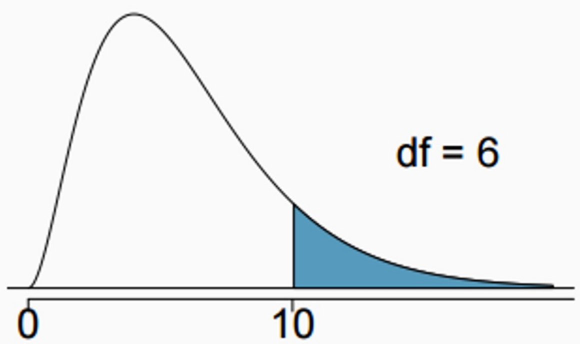
Estimate the shaded area under the chi-square curve with *df = 6*.



Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

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

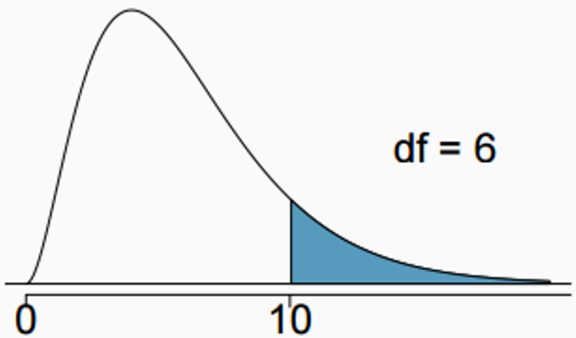
Estimate the shaded area under the chi-square curve with *df = 6*.



Upper tail	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df 1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

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

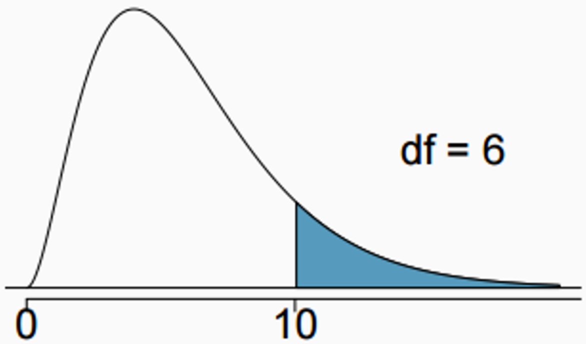
Estimate the shaded area under the chi-square curve with *df = 6*.



Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	*8.56*	*10.64*	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

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

Estimate the shaded area under the chi-square curve with *df = 6*.

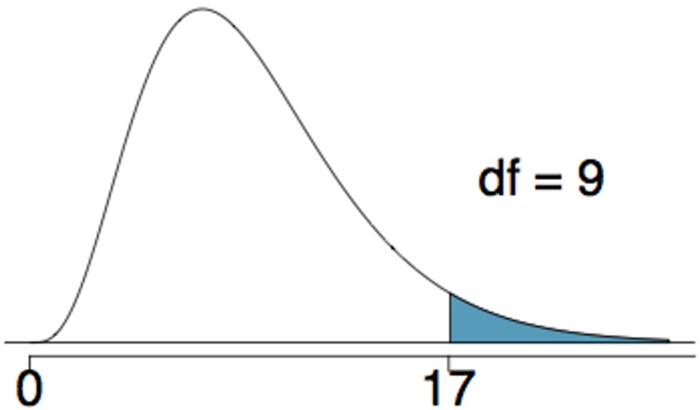


$P(\chi^2_{df=6} > 10)$
is between 0.1 and 0.2

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

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

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

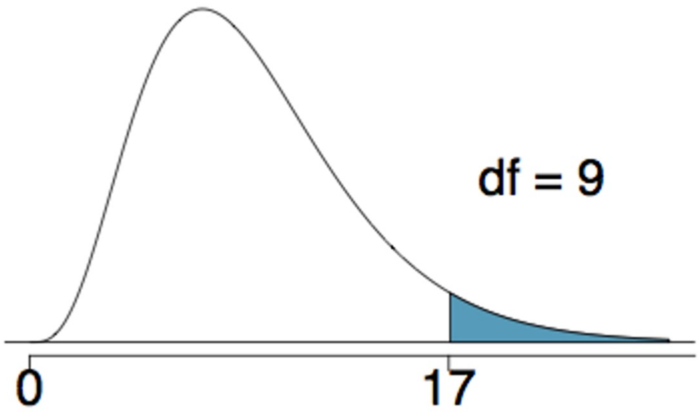


(a) between 0.01 and 0.02
(b) 0.02
(c) between 0.02 and 0.05
(d) 0.05
(e) between 0.05 and 0.10

Upper tail	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df 7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26

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

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



(a) between 0.01 and 0.02

(b) 0.02

(c) between 0.02 and 0.05

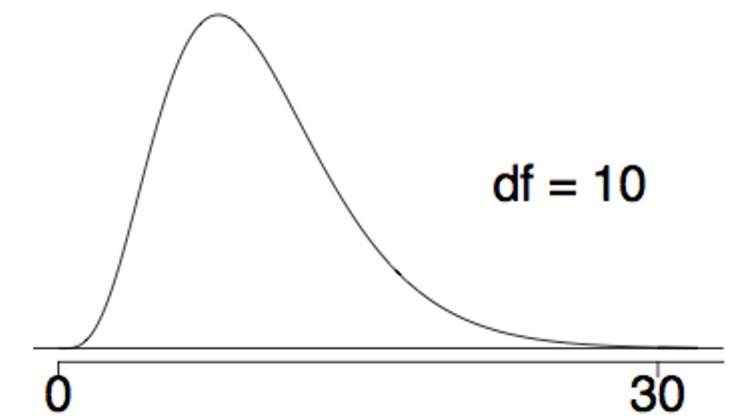
(d) 0.05

(e) between 0.05 and 0.10

Upper tail	0.3	0.2	0.1	*0.05*	*0.02*	0.01	0.005	0.001
df 7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
9	10.66	12.24	14.68	*16.92*	*19.68*	21.67	23.59	27.88
10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26

Finding areas under the chi-square curve (one more)

Estimate the shaded area (above 30) under the χ^2 curve with $df = 10$.

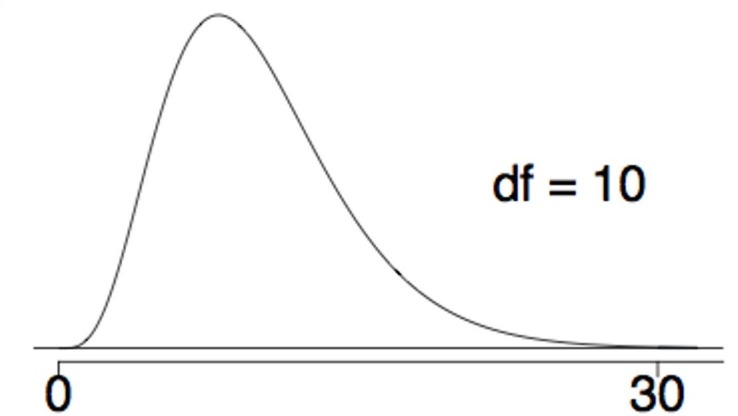


(a) between 0.005 and 0.001
(b) less than 0.001
(c) greater than 0.001
(d) greater than 0.3
(e) cannot tell using this table

Upper tail	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df 7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26

Finding areas under the chi-square curve (one more)

Estimate the shaded area (above 30) under the χ^2 curve with *df* = 10.



(a) greater than 0.3

(b) between 0.005 and 0.001

(c) less than 0.001

(d) greater than 0.001

(e) cannot tell using this table

Upper tail	0.3	0.2	0.1	0.05	0.02	0.01	0.005	*0.001*
df 7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	*29.59*
11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26

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.

$$df = k - 1$$

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.

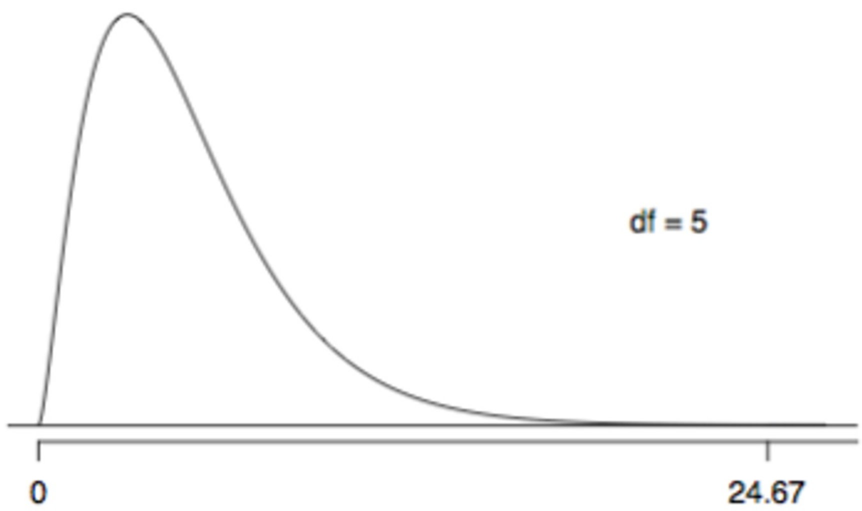
$$df = k - 1$$

- For our experiment, 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

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.

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.
- 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.

The χ^2 test

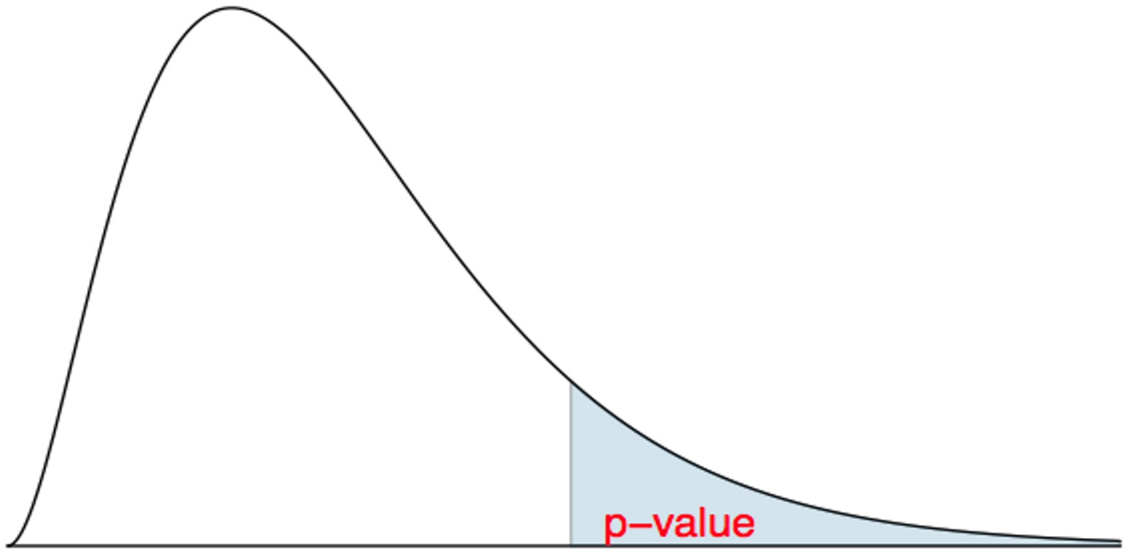
- Assume that you have a large population of items of k different types, and let p_i denote the probability of an item selected at random will be of type $i = 1, \dots, k$

- Let $p_1^o, \dots, p_k^o$ be numbers such that $p_i^o > 0$ $\sum p_i^o = 1$
- We want to test the hypothesis:
  - H_0: $p_i = p_i^o \; \forall \, i$ vs
  - H_1: $p_i \neq p_i^o$ *for at least one* i
- *Assume we have a data set of n observations, and* N_i *is the number of observations of type* i.
- *The expected number of observations of type* i *under the null hypothesis is* np_i^0
- Define the statistic $X^2 = \sum_{i=1}^{k} \frac{(N_i - np_i^o)^2}{np_i^0}$
- *Under the null, when* $n \to \infty, X^2 \sim \chi^2$ *with k-1 degrees of freedom.*

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

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 scenario (i.e. cell) must have at least 5 *expected* cases.
3. *df > 1*: Degrees of freedom must be greater than 1.

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

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%

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

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

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

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

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(30 - 13)^2}{13} = 22.23$$

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

Calculation of the test statistic

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27

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

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 <u>not</u> 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.

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.

Example: Independence

- You have a population of 520 people
    - 160/520 smoke.
    - 210/520 have CVD.

		CVD: Y	CVD: N	Total
Smoking	Y	120	40	160
	N	90	270	360
	Total	210	310	520

Contingency table

Example: Independence

Null Hypothesis ($\mathbf{H}_0$) : Smoking is independent of CVD

Alternative Hypothesis ($\mathbf{H}_1$) : Smoking is dependent of CVD

Mathematically:

$$\mathbf{H}_0 = \forall\, i,j \;\; p_{ij} = p_{i.} \times p_{.j}$$

$$\mathbf{H}_1 = \exists\, i,j\!: \;\; p_{ij} \neq p_{i.} \times p_{.j}$$

	CVD=0	CVD=1	
S=0	p_{00}	p_{01}	$p_{0.}$
S=1	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	1

$$p_{ij} = P(X = i,\; Y = j)$$

$$p_{i.} = P(X = i)$$

$$p_{.j} = P(Y = j)$$

Reminder: Independence:

$$\forall x, y\; P(\mathrm{Y} = \mathrm{y}, \mathrm{X} = \mathrm{x}) = P(Y = \mathrm{y})\mathrm{P}(\mathrm{X} = \mathrm{x})$$

Statistical Dependence

Smoking \ CVD	Y	N	Total
Y	120	40	160
N	90	270	360
Total	210	310	520

Contingency table

Smoking \ CVD	Y	N	Total
Y	.75	.25	1
N	.25	.75	1

Conditional Probability Distribution
$P(CVD|Smoking)$

Smoking \ CVD	Y	N	Total
Y	.2308	.0769	.3077
N	.1731	.5192	.6923
Total	.4038	.5962	1

Joint Probability Distribution
$P(CVD, Smoking)$

Smoking \ CVD	Y	N
Y	.5714	.1290
N	.4286	.8710
Total	1	1

Conditional Probability Distribution
$P(Smoking|CVD)$

Statistical Dependence

Smoking \ CVD	Y	N	Total
Y	120	40	160
N	90	270	360
Total	210	310	520

Contingency table

Smoking \ CVD	Y	N	Total
Y	.75	.25	1
N	.25	.75	1

Conditional Probability Distribution
$P(CVD|Smoking)$

P(Smoking)

Smoking \ CVD	Y	N	Total
Y	.2308	.0769	.3077
N	.1731	.5192	.6923
Total	.4038	.5962	1

Joint Probability Distribution
$P(CVD, Smoking)$

P(Smoking|CVD=yes)

Smoking \ CVD	Y	N
Y	.5714	.1290
N	.4286	.8710
Total	1	1

Conditional Probability Distribution
$P(Smoking|CVD)$

P(Smoking)≠P(Smoking|CVD=yes)

Test statistic: Expected counts

		CVD Y	CVD N	Total
Smoking	Y	.2308	.0769	.3077
	N	.1731	.5192	.6923
	Total	.4038	.5962	1

in your data

		CVD Y	CVD N	Total
Smoking	Y			.3077
	N			.6923
	Total	.4038	.5962	1

If Smoking and CVD were independent?

Are Smoking and CVD independent?

Smoking \ CVD	Y	N	Total
Y	.2308	.0769	.3077
N	.1731	.5192	.6923
Total	.4038	.5962	1

in your data

Smoking \ CVD	Y	N	Total
Y			.3077
N			.6923
Total	.4038	.5962	1

If Smoking and CVD were independent?

$$\boxed{P(Smoking = Yes, CVD = Yes) = P(Smoking = Yes) * P(CVD = Yes)}$$

Are Smoking and CVD independent?

		CVD		
		Y	N	Total
Smoking	Y	.2308	.0769	.3077
	N	.1731	.5192	.6923
	Total	.4038	.5962	1

in your data

		CVD		
		Y	N	Total
Smoking	Y			.3077
	N			.6923
	Total	.4038	.5962	1

If Smoking and CVD were independent?

$$\boxed{P(Smoking = Yes, CVD = Yes) = P(Smoking = Yes) * P(CVD = Yes) = 0.4038 * 0.3077}$$

Are Smoking and CVD independent?

Smoking \ CVD	Y	N	Total
Y	.2308	.0769	.3077
N	.1731	.5192	.6923
Total	.4038	.5962	1

in your sample

Smoking \ CVD	Y	N	Total
Y	.1242	.1835	.3077
N	.2796	.4127	.6923
Total	.4038	.5962	1

If Smoking and CVD were independent?

Are Smoking and CVD independent?

		CVD Y	CVD N
Smoking	Y	120	40
	N	90	270

counts in your data

		CVD Y	CVD N
Smoking	Y	65	95
	N	145	215

Expected counts If Smoking and CVD were independent

$$\boxed{P(Smoking = Yes, CVD = Yes) * \#\ samples = .1242*520}$$

- n_{ij}: Counts in your data (# observations in cell i,j)
- e_{ij}: Expected counts under H_0

$$X^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

What is the probability of observing a value *t* at least as extreme as the one you observed in your data?

p-value: $P\left(X^2 > x_{obs}^2 \middle| \mathrm{H}_0\right)$

df are the degrees of freedom, i.e. the number of parameters that are free to vary
For testing X ⫫ Y

$df = $ (# possible values of $X - 1$)×
(# of possible values of $Y - 1$)

in our example $df = (2-1)\times(2-1) = 1$