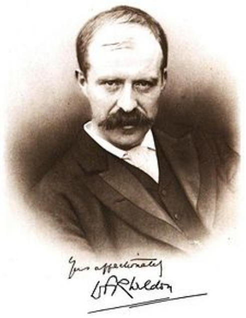
Lecture 25
Testing Goodness of Fit Using Chi-Square

Yibi Huang
Department of Statistics
University of Chicago

Weldon's dice

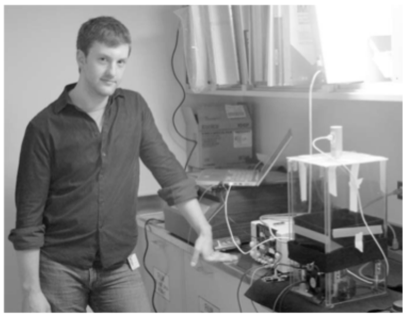
- Walter Frank Raphael Weldon (1860 - 1906), was an English evolutionary biologist and a founder of biometry. He was the joint founding editor of Biometrika, with Francis Galton and Karl Pearson.
- In 1894, he rolled 12 dice 26,306 times, and recorded the number of 5s or 6s (which he considered to be a success).
  - It was observed that 5s or 6s occurred more often than expected, and Pearson hypothesized that this was probably due to the construction of the dice. Most inexpensive dice have hollowed-out pips, and since opposite sides add to 7, the face with 6 pips is lighter than its opposing face, which has only 1 pip.



Labby's dice

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

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

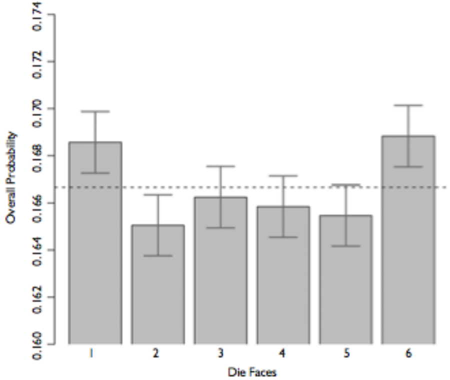


- The rolling-imaging process took about 20 seconds per roll.
  - Each day there were ∼150 images to process manually.
  - At this rate Weldon's experiment was repeated in a little more than six full days.
  - Recommended reading:

    https://galton.uchicago.edu/about/docs/2009/2009_dice_zac_labby.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, we expect each of the 6 faces to come up $\frac{12 \times 26,306}{6}$ times.

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

Hypotheses

Do these data provide convincing evidence that the 6 faces were not equally likely to come up?

H_0: The 6 faces of the die were equally likely to come up.

H_A: The 6 faces of the die were NOT equally likely to come up

The more deviant the observed counts from the expected counts under H_0, the stronger the evidence in favor of H_A

- How to measure how deviant the observed counts from the expected counts?

Pearson's Chi-Square Statistic

As an overall measure of the distance between the data and the expectations of the model, Pearson proposed the following χ^2-statistic

$$\begin{array}{c}\text{Pearson'}\\ \chi^2\text{-statistic}\end{array} = \sum \frac{(\text{obs'd count} - \text{exp'd count})^2}{\text{exp'd count}}$$

The more the observed frequencies deviate from the expected frequencies,

- the larger is the χ^2-statistic, and
- the stronger is the evidence against the fairness of the die.

Calculating the Chi-Square Statistic

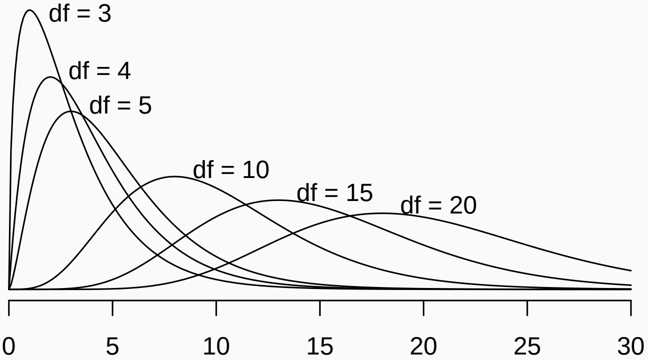
Outcome	Observed	Expected	$\frac{(O-E)^2}{E}$
1	53,222	52,612	$\frac{(53{,}222-52{,}612)^2}{52{,}612} = 7.07$
2	52,118	52,612	$\frac{(52{,}118-52{,}612)^2}{52{,}612} = 4.64$
3	52,465	52,612	$\frac{(52{,}465-52{,}612)^2}{52{,}612} = 0.41$
4	52,338	52,612	$\frac{(52{,}338-52{,}612)^2}{52{,}612} = 1.43$
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

So the χ^2-statistic for Labby's experiment is 24.74.

Is this number big or small?

We need to know the sampling distribution of the χ^2-statistic.

The Chi-Square (χ^2) Distribution



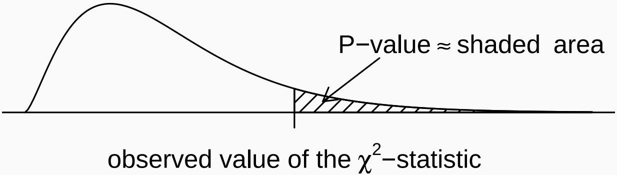
- There is one curve with each number of **degree of freedom**
- All χ^2-curves are right-skewed
- As the degrees of freedom ↑, the curves flatten out and move off to the right, and become less skewed (more symmetric)
- Expected value $=$ *df*, SD $= \sqrt{df}$

Approximate Distribution of the χ^2-Statistic

If the number of observations (sample size) is large, the χ^2-statistics has an *approximate χ^2-distribution with $k-1$ degrees of freedom*. Here

$$k = \text{\# of summands in the } \chi^2\text{-statistic}$$

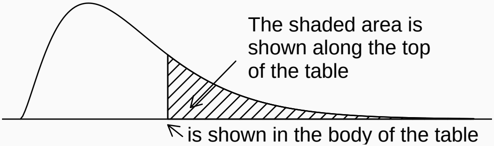
- p-value = upper tail area under the χ^2 curve



- For Labby's experiment, there are __5__ degrees of freedom.
- Rule of thumb for sample size required: *all expected counts should be ≥ 5.*

Chi-Square Probability Table (p.432 in text)

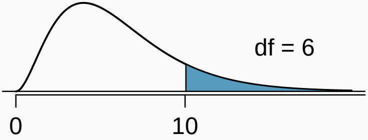
The χ^2-curve, with degrees of freedom shown along the left of the table.



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
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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Practice: Finding p-Value for a χ^2-Statistic

Suppose a χ^2-statistic is 10.3, with df = 6. Find the p-value.



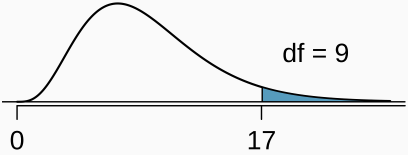
p-value = $P(\chi^2_{df=6} > 10.3)$
is between 0.1 and 0.2

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

```
> pchisq(10.3, df = 6, lower.tail = FALSE)
[1] 0.1125737
```

Practice: Finding p-Value for a χ^2-Statistic

Suppose a χ^2-statistic is 17.56, with df = 9. Find the p-value.



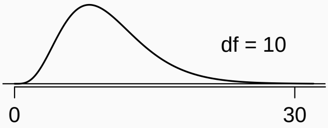
p-value = $P(\chi^2_{df=9} > 17.56)$
is between 0.02 and 0.05

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

```
> pchisq(17.56, df = 9, lower.tail = FALSE)
[1] 0.04063539
```

Practice: Finding p-Value for a χ^2-Statistic

Suppose a χ^2-statistic is 30.9, with df = 10. Find the p-value



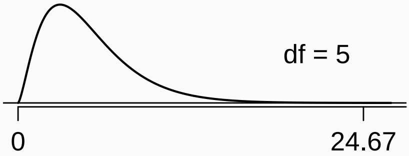
p-value = $P(\chi^2_{df=10} > 30.9)$
is less than 0.001

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001	
df	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	

```
> pchisq(30.9, df = 10, lower.tail = FALSE)
[1] 0.0006094554
```

Back to Labby's Dice

The χ^2-statistic for Labby's experiment is 24.67, with df $= 6 - 1 = 5$.



p-value = $P(\chi^2_{df=5} > 24.67)$
is less than 0.001
(By R, p-value $= 0.00016$)

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*	→

Conclusion:

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: Chi-square Test for Goodness of Fit

Suppose we have a hypothetical model about the distribution of a categorical variable

Category	1	2	$\cdots$	k
Probability	p_1	p_2	$\cdots$	p_k

Then we collect data

Category	1	2	$\cdots$	k
Observed Counts	O_1	O_2	$\cdots$	O_k

and want to test whether the data are i.i.d. observations from the hypothesized distribution.

H_0: the data are i.i.d. observations from the hypothesized distribution.

H_A: the data are NOT i.i.d. observations from the hypothesized distribution.

Recap: Chi-square Test for Goodness of Fit (Cont'd)

When H_0 is true, the *expected counts* for the ith category is np_i where n is the total number of observations.

category	1	2	$\cdots$	k	Total
observed counts	O_1	O_2	$\cdots$	O_k	n
expected counts	np_1	np_2	$\cdots$	np_k	n

- The expected counts np_i need NOT to be a whole number. Do not round it!

The χ^2-statistic is

$$\chi^2 = \sum_i \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} = \sum_i \frac{(O_i - np_i)^2}{np_i}$$

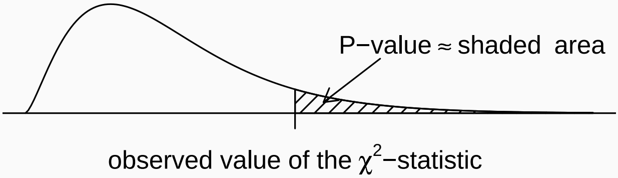
and the degrees of freedom $= k - 1$ (one less than the number of categories).

Recap: Chi-square Test for Goodness of Fit (Cont'd)

Why $k-1$ degrees of freedom?

- Because there are $k-1$ parameters: $p_1, p_2, \ldots, p_k$ with the constraint $p_1 + p_2 + \ldots + p_k = 1$.

The *p-value* is approx. the upper tail area under the χ^2 curve with $k-1$ degrees of freedom



- The chi-square approximation works well when all expected counts are at least 5.

Example: Mendel's Genetic Model

The International Rice Research Institute in the Philippines develops new lines of rice which combine high yields with resistance to disease and insects. The technique involves crossing different lines to get a new line which has the most advantageous combination of genes. Detailed genetic modeling is required. One project involved breeding new lines to resist the "brown plant hopper" (an insect): 374 lines were raised, with the results shown below.

	Number of lines	Model
All plants resistant	97	0.25
Mixed: some plants resistant, some susceptible	184	0.5
All plants susceptible	93	0.5

According to the IRRI model, the lines are independent: each line has a 25% chance to be resistant, a 50% to be mixed, and a 25% chance to be susceptible. Are the data consistent with this model?

Example: Mendel's Genetic Model

According to the IRRI model, the expected count of for each type is

Type	Model	observed count	expected count	(Obs-Exp)^2/Obs
Resistent	25%	97	$374 \times 0.25 = 93.5$	$\frac{(97-93.5)^2}{93.5} \approx 0.1310$
Mixed	50%	184	$374 \times 0.50 = 187$	$\frac{(184-187)^2}{187} \approx 0.0481$
Susceptible	25%	93	$374 \times 0.25 = 93.5$	$\frac{(93-93.5)^2}{93.5} \approx 0.0027$
Total	100%	374	374	χ^2-statistics $= 0.1818$

The df is $3 - 1 = 2$. P-value ≥ 0.3 because the χ^2-statistics $= 0.1818 < 2.41$, showing the consistency of the results with the model.

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

Example: Seasonal Variation of Suicide Rate

Suicide Counts in US by month in 1970

Month	# of suicides	days/ month	expected counts
Jan	1867	31	2021.889
Feb	1789	28	1826.222
Mar	1944	31	2021.889
Apr	2094	30	1956.667
May	2097	31	2021.889
Jun	1981	30	1956.667
July	1887	31	2021.889
Aug	2024	31	2021.889
Sept	1928	30	1956.667
Oct	2032	31	2021.889
Nov	1978	30	1956.667
Dec	1859	31	2021.889
Total	23480	365	23480

Does the suicide rate vary seasonally, or is it constant from day to day?

If the suicide rate is constant from day to day, the chance that a suicide occurs in January is 31/365. The expected number of suicides in January is thus

$$\text{(total number of suicides)} \times \frac{31}{365} = 2021.889.$$

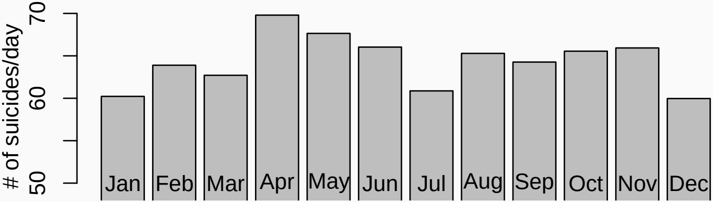
Source: The National Center for Health Statistics (1970)

Example: Seasonal Variation of Suicide Rate (Cont'd)

- The χ^2-statistic is

$$\frac{(1867-2021.889)^2}{2021.889}+\frac{(1789-1826.222)^2}{1826.222}+\ldots+\frac{(1859-2021.889)^2}{2021.889}=51.18$$

  with $12-1=11$ degrees of freedom.
- *p*-value is less than 0.001



https://en.wikipedia.org/wiki/Seasonal_effects_on_suicide_rates

When there are only two categories, the χ^2-test is equivalent to a two-sided one-sample test of proportion $H_0 : p = p_0$.

Category	Success	Failure	Total
probability under H_0	p_0	$1 - p_0$	1
observed counts	X	$n - X$	n
expected counts	np_0	$n(1 - p_0)$	n

$$
\begin{aligned}
\chi^2 = \sum \frac{(O-E)^2}{E} &= \frac{(X - np_0)^2}{np_0} + \frac{(n - X - n(1-p_0))^2}{n(1-p_0)} \\
&= \frac{(X - np_0)^2}{np_0} + \frac{(X - np_0)^2}{n(1-p_0)} \\
&= \frac{(X - np_0)^2}{n}\left(\frac{1}{p_0} + \frac{1}{1-p_0}\right) \\
&= \frac{(X - np_0)^2}{n}\left(\frac{p_0 + (1-p_0)}{p_0(1-p_0)}\right) = \frac{(X - np_0)^2}{np_0(1-p_0)} \\
&= \left(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\right)^2 \quad \text{where } \hat{p} = \frac{X}{n}
\end{aligned}
$$

So the chi-square statistic is simply the square of z-statistic $= \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$.

Furthermore, the chi-square distribution with df = 1 is simply the square of $N(0, 1)$.

So the two tests give identical p-values.