# STA221

Neil Montgomery

Last edited: 2017-01-17 10:53

goodness-of-fit testing ("Comparing Counts", Ch. 23 of textbook)

# detour 1 - what tends to have a Normal distribution?

In Stats 101 you will have encountered (at least) Binomial and Normal distribuions.

# detour 1 - what tends to have a Normal distribution?

In Stats 101 you will have encountered (at least) Binomial and Normal distribuions.

And we've learned that $\overline{X}$ is always approximately Normal when the sample is large enough.

Sums of random things also tend to be approximately Normal.

# detour 1 - what tends to have a Normal distribution?

In Stats 101 you will have encountered (at least) Binomial and Normal distribuions.

And we've learned that $\overline{X}$ is always approximately Normal when the sample is large enough.

Sums of random things also tend to be approximately Normal.

A special case is the so-called "Normal approximation to the Binomial".

# detour 1 - what tends to have a Normal distribution?

In Stats 101 you will have encountered (at least) Binomial and Normal distribuions.

And we've learned that $\overline{X}$ is always approximately Normal when the sample is large enough.

Sums of random things also tend to be approximately Normal.

A special case is the so-called "Normal approximation to the Binomial".

Why? Recall that a Binomial probability model is used to *count* the number of "*successes*" in $n$ "trials".

# detour 1 - what tends to have a Normal distribution?

In Stats 101 you will have encountered (at least) Binomial and Normal distribuions.

And we've learned that $\overline{X}$ is always approximately Normal when the sample is large enough.

Sums of random things also tend to be approximately Normal.

A special case is the so-called "Normal approximation to the Binomial".

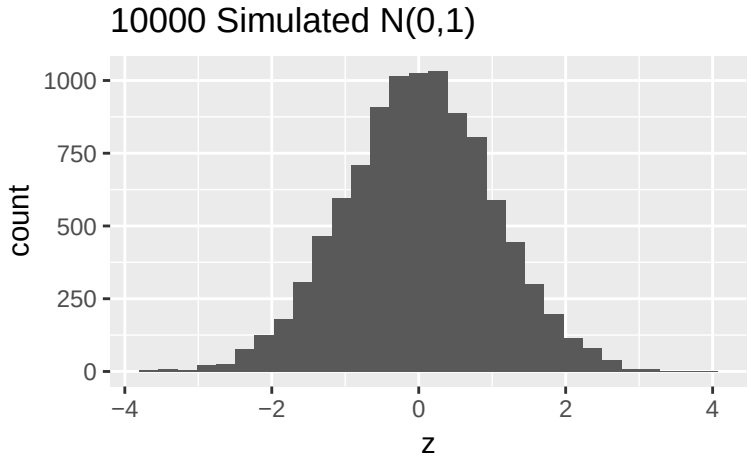
Why? Recall that a Binomial probability model is used to *count* the number of "*successes*" in $n$ "trials".

Let's map "success" to the number 1 and "failure" to the number 0.

**Counting 1s in a sequence of 0s and 1s is exactly equivalent to adding up all the 0s and 1s**

# detour 2.1 - what happens when you look at the square of a normal?

My compute can simulate random "draws" from a standard normal (N(0,1)) distribution, resulting in a histogram such as:

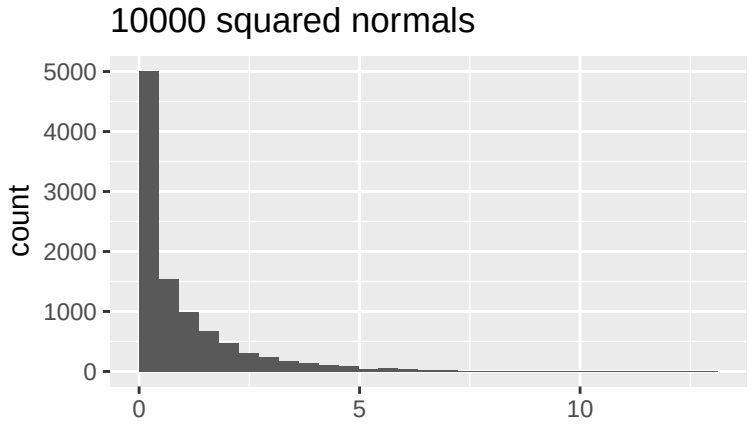


# detour 2.2 - what what happens when you look at the square of a normal?

I could take all of those simulated standard normals and square them, and make a histogram of the result, which would give:

# detour 2.2 - what what happens when you look at the square of a normal?

I could take all of those simulated standard normals and square them, and make a histogram of the result, which would give:

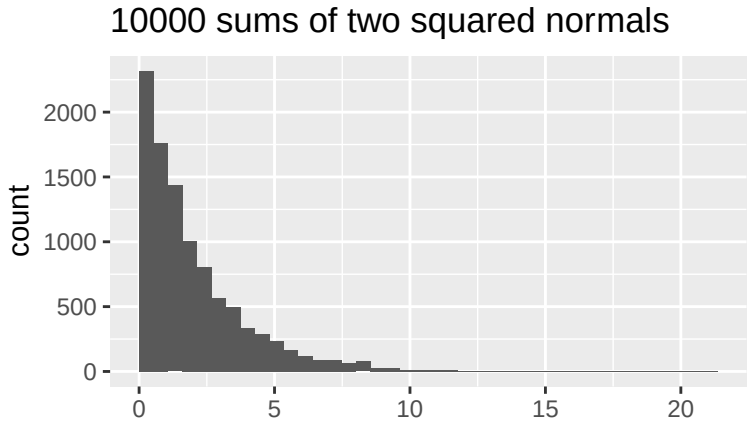


# detour 2.3 - sum of squared normals?

I can simulate *two* columns of standard normals, square them *both*, add the results, and make a histogram of the result:

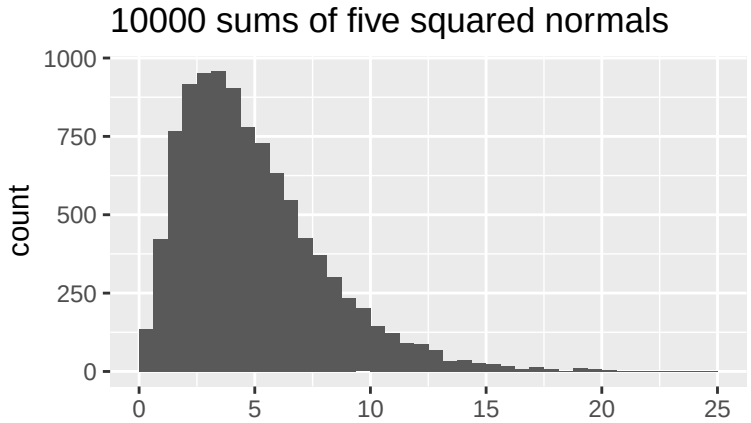
# detour 2.3 - sum of squared normals?

I can simulate *two* columns of standard normals, square them *both*, add the results, and make a histogram of the result:



# detour 2.4 - sum of many squared normals?

I can make several columns of normals, square them, add them up, and make a histogram. Here's the histogram with 5 columns of normals:



# detour - the $\chi^2$ family of distributions

If you have $n$ independent standard normals, the sum of their squares will have a $\chi^2_n$ distribution.

# detour - the $\chi^2$ family of distributions

If you have $n$ independent standard normals, the sum of their squares will have a $\chi^2_n$ distribution.

The $n$ is a parameter going by the name "degrees of freedom."

# detour - the $\chi^2$ family of distributions

If you have $n$ independent standard normals, the sum of their squares will have a $\chi^2_n$ distribution.

The $n$ is a parameter going by the name "degrees of freedom."

If you have $n$ general $N(\mu, \sigma)$, say called $X_1, X_2, \ldots, X_n$, you could *standardize them*:

$$Z_i = \frac{X_i - \mu}{\sigma}$$

and then the sums of the squares of these $Z_i$ will have a $\chi^2_n$ distribution.

# detour - the $\chi^2$ family of distributions

If you have $n$ independent standard normals, the sum of their squares will have a $\chi^2_n$ distribution.

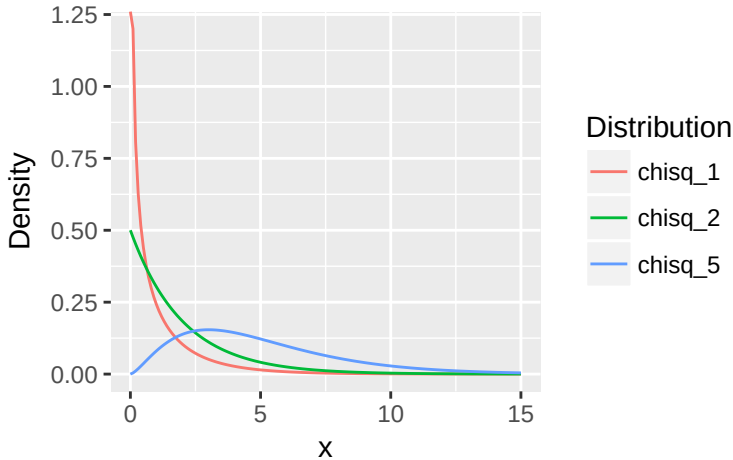
The $n$ is a parameter going by the name "degrees of freedom."

If you have $n$ general $N(\mu, \sigma)$, say called $X_1, X_2, \ldots, X_n$, you could *standardize them*:

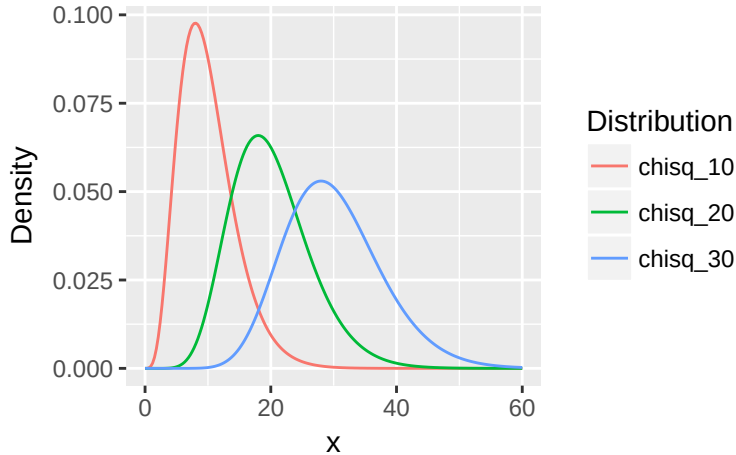
$$Z_i = \frac{X_i - \mu}{\sigma}$$

and then the sums of the squares of these $Z_i$ will have a $\chi^2_n$ distribution.

# detour - pictures of some $\chi^2_n$ distributions



# detour - pictures of more $\chi^2_n$ distributions




Note: the average of a $\chi^2_n$ distribution is just $n$.

## ever wonder why the sample variance is divided by $n-1$?

Look at the formula for sample variance:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

The numerator is a sum of $n$ squares, but the denominator is $n-1$. Why?

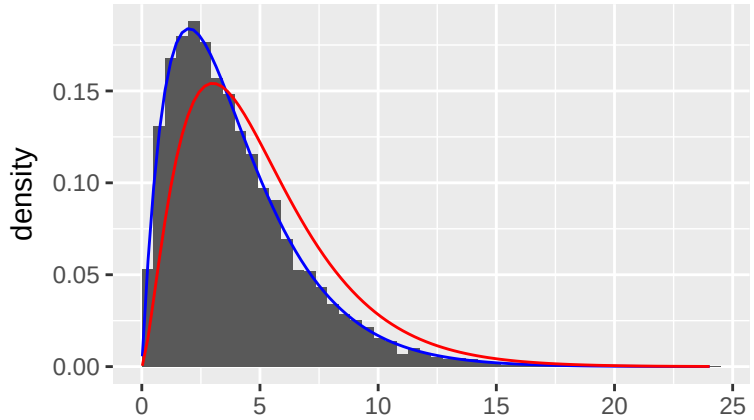
## pictures of $\sum_{i=1}^{5}(x_i - \overline{x})^2$

I can simulate samples of size, say, 5 and compute that numerator, and make a histogram.

## pictures of $\sum_{i=1}^{5}(x_i - \overline{x})^2$

I can simulate samples of size, say, 5 and compute that numerator, and make a histogram.

Here it is, with the $\chi^2_4$ distribution in blue and the $\chi^2_5$ in red:



# a heuristic explanation

$s^2$ is calculated after fixing the value of $\overline{x}$

# a heuristic explanation

$s^2$ is calculated after fixing the value of $\overline{x}$

So given $\overline{x}$ and *any* $n-1$ of the $n$ raw values, I can calculate that other raw value.

# a heuristic explanation

$s^2$ is calculated after fixing the value of $\overline{x}$

So given $\overline{x}$ and *any* $n-1$ of the $n$ raw values, I can calculate that other raw value.

We say $s^2$ (given $\overline{x}$) only has $n-1$ degrees of freedom.

is there evidence that something doesn't follow a given distribution?

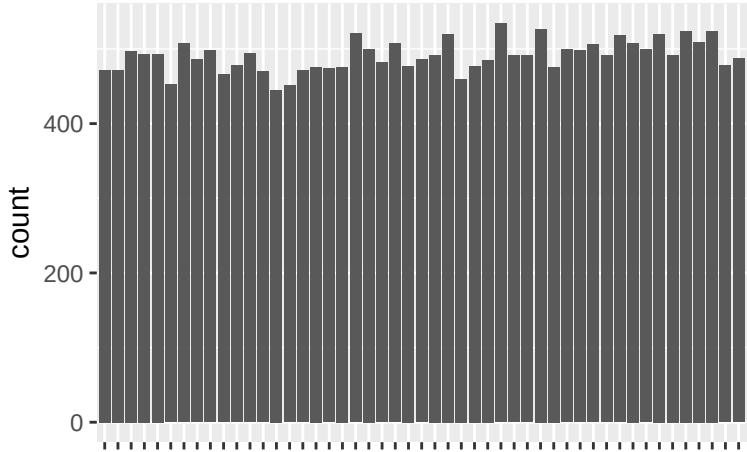
# is a lottery "fair"

Lotto 6/49 is a Canadian lottery in which 49 identical balls are mixed together and 7 are selected, now twice per week. People can win money based on how many of the numbers they have out of the 6 on their ticket.

I found a list of every number ever picked here:
`http://portalseven.com/lottery/canada_lotto_649.jsp`

```
## # A tibble: 3,437 × 8
##                  date  num1  num2  num3  num4  num5  num6 bonus
##                 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Sat, Jan 14, 2017     1     6    19    30    32    44    33
## 2 Wed, Jan 11, 2017    24    34    36    38    42    43    30
## 3  Sat, Jan 7, 2017     1    10    18    19    23    27    48
## 4  Wed, Jan 4, 2017     2    11    13    23    35    48    30
## 5 Sat, Dec 31, 2016     3     5    14    18    26    28    40
## # ... with 3,432 more rows
```

all 49 numbers should appear with roughly the same frequency



# categorical data, cells, observed cell counts

The dataset (now) consists of one variable called `numbers`. This is a *categorical*, or *factor* variable with 49 possible *levels*. There are 24050 observations.

# categorical data, cells, observed cell counts

The dataset (now) consists of one variable called `numbers`. This is a *categorical*, or *factor* variable with 49 possible *levels*. There are 24050 observations.

A categorical variable is summarized by producing a table of *observed cell counts* (notation: $O_i$). In this case:

```
## # A tibble: 49 × 2
##   numbers   O_i
##    <fctr> <int>
## 1       1   472
## 2       2   472
## 3       3   497
## 4       4   493
## 5       5   493
## # ... with 44 more rows
```

# expected cell counts

If Lotto 6/49 is actually fair, each number would appear with probability $1/49 = 0.0204$ each.

After 24050 numbers have been selected, we would expect to see:

$$24050 \cdot \frac{1}{49} = 490.82$$

of each number.

These are called *expected cell counts* — calulated under the assumption of fairness as defined in this example. (Notation: $E_i$)

# measuring the deviation from the assumption of fairness

Each $O_i$ is a count (i.e. a sum of 0s and 1s), which will have an approximate normal distribution. It turns out:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

has a standard normal distribution, as long as there are enough 1s in the sample.

# measuring the deviation from the assumption of fairness

Each $O_i$ is a count (i.e. a sum of 0s and 1s), which will have an approximate normal distribution. It turns out:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

has a standard normal distribution, as long as there are enough 1s in the sample.

How many? As long as $E_i \geq 10$, the approximation will be good.

# measuring the deviation from the assumption of fairness

Each $O_i$ is a count (i.e. a sum of 0s and 1s), which will have an approximate normal distribution. It turns out:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

has a standard normal distribution, as long as there are enough 1s in the sample.

How many? As long as $E_i \geq 10$, the approximation will be good.

(Note: this is exactly like the $np \geq 10$ or $np \geq 5$ suggestion that is given for the accuracy of a normal approximation to a binomial.)

# measuring the deviation from the assumption of fairness

Each $O_i$ is a count (i.e. a sum of 0s and 1s), which will have an approximate normal distribution. It turns out:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

has a standard normal distribution, as long as there are enough 1s in the sample.

How many? As long as $E_i \geq 10$, the approximation will be good.

(Note: this is exactly like the $np \geq 10$ or $np \geq 5$ suggestion that is given for the accuracy of a normal approximation to a binomial.)

The overall deviation is measured as:

$$\sum_{i=1}^{n} \left( \frac{O_i - E_i}{\sqrt{E_i}} \right)^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

# measuring the deviation - compared to what?

The expected cell counts $E_i$ are computed *for a given fixed sample size* $N$.

# measuring the deviation - compared to what?

The expected cell counts $E_i$ are computed *for a given fixed sample size* $N$.

So given $n$ along with any of the $n-1$ expected cell counts, we could compute that other expected cell count.

# measuring the deviation - compared to what?

The expected cell counts $E_i$ are computed *for a given fixed sample size* $N$.

So given $n$ along with any of the $n-1$ expected cell counts, we could compute that other expected cell count.

(This might *seem* trivial in this example because all the expected cell counts are the same - but this is only because our hypothesis is that all the cell probabilities are the same.)

# measuring the deviation - compared to what?

The expected cell counts $E_i$ are computed *for a given fixed sample size* $N$.

So given $n$ along with any of the $n-1$ expected cell counts, we could compute that other expected cell count.

(This might *seem* trivial in this example because all the expected cell counts are the same - but this is only because our hypothesis is that all the cell probabilities are the same.)

We say

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

has $n-1$ degrees of freedom, and it follows (approximately) a $\chi^2_{n-1}$ distribution.

# let's measure the deviation

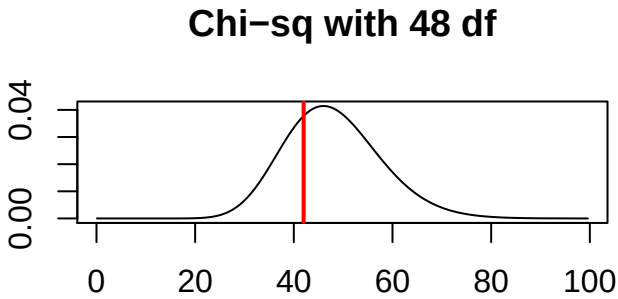
Here are the first few deviations (with $(O_i - E_i)^2/E_i$ called `D_i` for short):

```
## # A tibble: 49 × 4
##   numbers   O_i    E_i         D_i
##    <fctr> <int>  <dbl>       <dbl>
## 1       1   472 490.82 0.721634000
## 2       2   472 490.82 0.721634000
## 3       3   497 490.82 0.077813455
## 4       4   493 490.82 0.009682572
## 5       5   493 490.82 0.009682572
## # ... with 44 more rows
```

The sum of the `D_i` column is 41.99. Is this number surprising?

## surprising, compared to what?

We know we should compare this number with the $\chi^2_{48}$ distribution. Here we can see we are not surprised. There is no evidence that Lotto 6/49 is unfair.



# goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

# goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

The null hypothesis is the result of statement that data follows a certain distribtuion of interest.

# goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

The null hypothesis is the result of statement that data follows a certain distribtuion of interest.

In the Lotto example, technically this statement is:

$$H_0 : p_1 = p_2 = \cdots = p_{49} = \frac{1}{49}$$

But usually we just make $H_0$ a simple written statement:

$$H_0 : \text{ the probabilities are all the same.}$$

# goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

The null hypothesis is the result of statement that data follows a certain distribtuion of interest.

In the Lotto example, technically this statement is:

$$H_0 : p_1 = p_2 = \cdots = p_{49} = \frac{1}{49}$$

But usually we just make $H_0$ a simple written statement:

$$H_0 : \text{ the probabilities are all the same.}$$

The "distribution of interest" is technically the "discrete uniform distribution on the outcomes $\{1, 2, 3, \ldots, 49\}$"

# goodness of fit as formal hypothesis test - I

Goodness-of-fit testing is a form of hypothesis testing.

The null hypothesis is the result of statement that data follows a certain distribtuion of interest.

In the Lotto example, technically this statement is:

$$H_0 : p_1 = p_2 = \cdots = p_{49} = \frac{1}{49}$$

But usually we just make $H_0$ a simple written statement:

$$H_0 : \text{ the probabilities are all the same.}$$

The "distribution of interest" is technically the "discrete uniform distribution on the outcomes $\{1, 2, 3, \ldots, 49\}$"

The alternative hypothesis is the negation of the null. We don't normally bother to write it down.

# goodness of fit as formal hypothesis test - II

Given a sample size $N$ and the null hypothesis probabilities, compute the $n$ expected cell counts. In this case:

$$E_i = Np_i = 24050 \cdot \frac{1}{49} = 490.82$$

# goodness of fit as formal hypothesis test - II

Given a sample size $N$ and the null hypothesis probabilities, compute the $n$ expected cell counts. In this case:

$$E_i = Np_i = 24050 \cdot \frac{1}{49} = 490.82$$

Compute the *observed value of the test statistic*:

$$\chi^2_{\text{obs}} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 41.99$$

# goodness of fit as formal hypothesis test - II

Given a sample size $N$ and the null hypothesis probabilities, compute the $n$ expected cell counts. In this case:

$$E_i = Np_i = 24050 \cdot \frac{1}{49} = 490.82$$

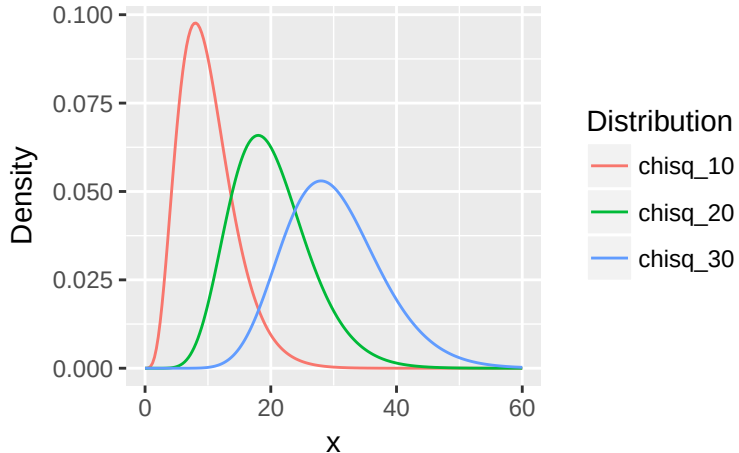
Compute the *observed value of the test statistic*:

$$\chi^2_{\text{obs}} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 41.99$$

Calculate the p-value based on $\chi^2_{\text{obs}}$ being approximately $\chi^2_{n-1}$.

# goodness-if-fit testing p-value

A p-value is the *probability of observing a more extreme value*, in the sense of being further from where the null hypothesis "lives", which is where in this case?



## goodness-of-fit testing p-value

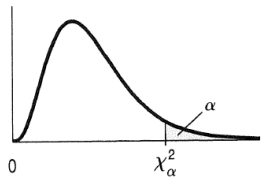
The p-value is $P(\chi^2_{48} \geq 41.99) = 0.7165747$

# goodness-of-fit testing p-value

The p-value is $P(\chi^2_{48} \geq 41.99) = 0.7165747$

On tests you'll need to use a table. Here's a close-up of the book's table:

**Table X**
Values of $\chi^2_\alpha$




| Right-tail probability | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 |
|---|---|---|---|---|---|
| df | | | | | |
| 1 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 |
| 2 | 4.605 | 5.991 | 7.378 | 9.210 | 10.597 |
| 3 | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 |
| 4 | 7.779 | 9.488 | 11.143 | 13.277 | 14.860 |
| 5 | 9.236 | 11.070 | 12.833 | 15.086 | 16.750 |
| 6 | 10.645 | 12.592 | 14.449 | 16.812 | 18.548 |
| 7 | 12.017 | 14.067 | 16.013 | 18.475 | 20.278 |
| 8 | 13.362 | 15.507 | 17.535 | 20.090 | 21.955 |
| 9 | 14.684 | 16.919 | 19.023 | 21.666 | 23.589 |
| 10 | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 |
| 11 | 17.275 | 19.675 | 21.920 | 24.725 | 26.757 |
| 12 | 18.549 | 21.026 | 23.337 | 26.217 | 28.300 |
| 13 | 19.812 | 22.362 | 24.736 | 27.688 | 29.819 |

# goodness-of-fit testing p-value (from table)

| | | | | | |
|---|---|---|---|---|---|
| 28 | 37.916 | 41.337 | 44.461 | 48.278 | 56.994 |
| 29 | 39.087 | 42.557 | 45.722 | 59.588 | 52.336 |
| 30 | 40.256 | 43.773 | 46.979 | 50.892 | 53.672 |
| 40 | 51.805 | 55.759 | 59.342 | 63.691 | 66.767 |
| 50 | 63.167 | 67.505 | 71.420 | 76.154 | 79.490 |
| 60 | 74.397 | 79.082 | 83.298 | 88.381 | 91.955 |
| 70 | 85.527 | 90.531 | 95.023 | 100.424 | 104.213 |

From a table the best you can do is to estimate the p-value.

***All this together is called the "$\chi^2$ goodness-of-fit test."***

applications of $\chi^2$ goodness-of-fit testing to two-way tables

# contingency tables

Recall the gas pipelines data:

```
## # A tibble: 1,000 × 4
##     Leak  Size Material Pressure
##   <fctr> <ord>   <fctr>   <fctr>
## 1     No  1.75  Aldyl A     High
## 2     No  1.75  Aldyl A      Med
## 3     No     1  Aldyl A      Low
## 4    Yes   1.5    Steel      Med
## 5     No     1    Steel     High
## # ... with 995 more rows
```

The (only?) suitable numerical summary for two categorical/factor variables at a time is a so-called contingency table, or two-way table.

# two-way table for “Leak” and “Pressure”

| | High | Low | Med | Sum |
|---|---|---|---|---|
| No | 277 | 278 | 247 | 802 |
| Yes | 71 | 66 | 61 | 198 |
| Sum | 348 | 344 | 308 | 1000 |

# the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

# the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

This question is answered using a *test of homogeneity*.

# the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

This question is answered using a *test of homogeneity*.

Question 2: are the rows and columns *independent*?

# the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

This question is answered using a *test of homogeneity*.

Question 2: are the rows and columns *independent*?

This question is answered using a *test of independence*.

# the two questions

Question 1: do the rows (or columns) come from the *same* distribution?

This question is answered using a *test of homogeneity*.

Question 2: are the rows and columns *independent*?

This question is answered using a *test of independence*.

***The mechanics of both tests are identical. Only the interpretation is (slightly) different.***

# two-way table again

Count version:

| | High | Low | Med | Sum |
|---|---|---|---|---|
| No | 277 | 278 | 247 | 802 |
| Yes | 71 | 66 | 61 | 198 |
| Sum | 348 | 344 | 308 | 1000 |

Proportion version:

| | High | Low | Med | Sum |
|---|---|---|---|---|
| No | 0.277 | 0.278 | 0.247 | 0.802 |
| Yes | 0.071 | 0.066 | 0.061 | 0.198 |
| Sum | 0.348 | 0.344 | 0.308 | 1.000 |