week 6 day 2

“Exact testing for model/data fit for log-linear models”
“Part One.”
“Algebraic & Geometric Methods in Statistics”

Sonja Petrović
Created for Math/Stat 561

Feb 15, 2023.

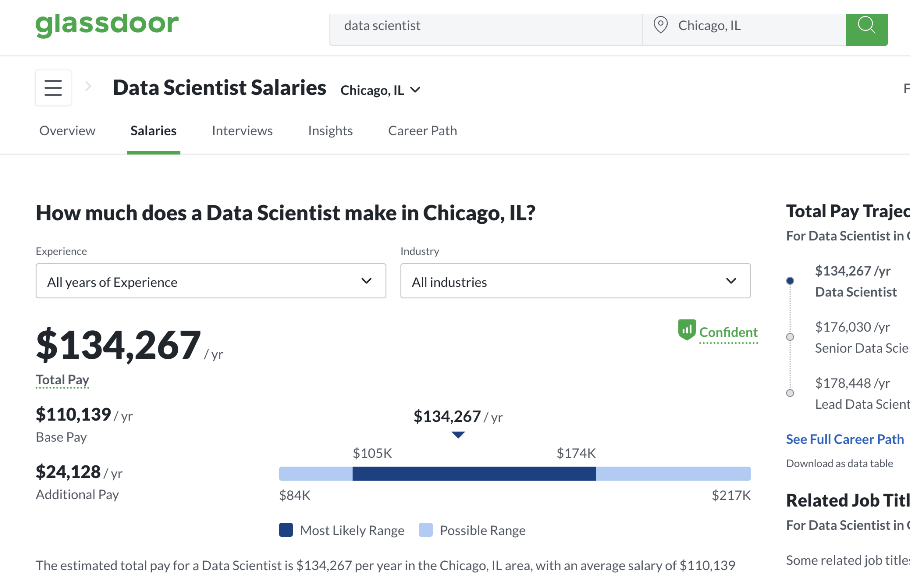
Agenda

- Chapter 9 from our textbook: Fisher's exact test
- Part of chapter 8, as we may need the cone of sufficient statistics.

Goals

- Understand hypotheses testing for model/data fit
- What is a p-value for a goodness-of-fit test?
- Asymptotic vs. exact tests
- Fisher's test and example
- General goodness of fit test for log-linear models
- Open problems and relation to projects!

A simple search: Chicago data science salary data



Data Scientist Salaries Chicago, IL

How much does a Data Scientist make in Chicago, IL?

Experience: All years of Experience

Industry: All industries

$134,267 / yr

Total Pay

Confident

$110,139 / yr

Base Pay

$24,128 / yr

Additional Pay

$134,267 / yr

$105K $174K

$84K $217K

■ Most Likely Range ■ Possible Range

The estimated total pay for a Data Scientist is $134,267 per year in the Chicago, IL area, with an average salary of $110,139

Total Pay Trajec

For Data Scientist in (

● $134,267 /yr
Data Scientist

○ $176,030 /yr
Senior Data Scie

○ $178,448 /yr
Lead Data Scient

See Full Career Path

Download as data table

Related Job Titl

For Data Scientist in (

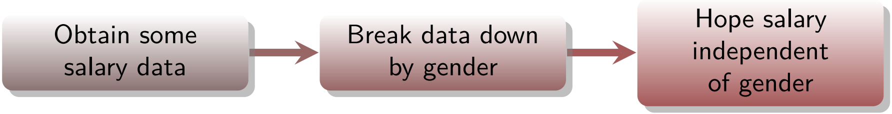
Some related job title

At the heart of statistical reasoning

- Given: data, find out if it is usual/expected? surprising/outlier? quantify??

At the heart of statistical reasoning

- Given: data, find out if it is usual/expected? surprising/outlier? quantify??
- Do all genders get fair salary in Chicago for data science jobs?



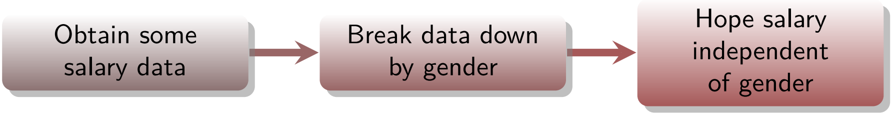
Obtain some salary data

Break data down by gender

Hope salary independent of gender

At the heart of statistical reasoning

- Given: data, find out if it is usual/expected? surprising/outlier? quantify??
- Do all genders get fair salary in Chicago for data science jobs?



Obtain some salary data

Break data down by gender

Hope salary independent of gender

- We expect a certain 'shape' of the data. A certain… distribution!

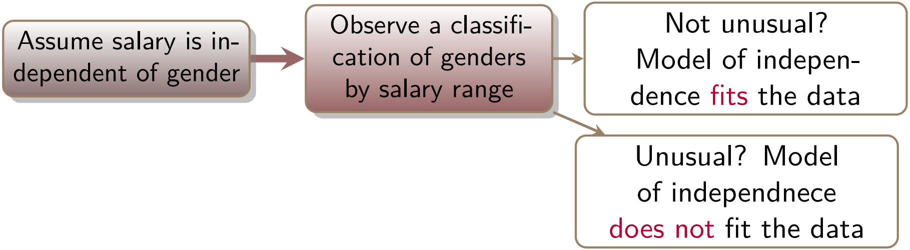
YOUR everyday intuition $\mapsto$ formal framework.

	M	F	T/Nb	totals
$\leq$ 135K	?	?	?	**13**
$>$ 135K	?	?	?	**13**
totals	**10**	**10**	**6**	**26**

Formal reasoning with data: independence example

- **Modeling**: Construct a statistical model for independence.
- **Question**: Does the model fit the observed set of gender vs. salary ranges?

  (Can it adequately explain how the salary data was generated?)
- **Process**:



Assume salary is independent of gender

Observe a classification of genders by salary range

Not unusual? Model of independence fits the data

Unusual? Model of independnece does not fit the data

What are tests of goodness of fit?

What familiar data structure is this:

```
        gender
range     M  F Nb
  <=135K  8  1  4
  > 135K  2  9  2
```

The *research hypothesis* can be made about a model: for example, since it looks like there's an association between lower salary bracket and gender, maybe the data is enough evidence to refute:

What are tests of goodness of fit?

What familiar data structure is this:

```
        gender
range      M  F Nb
  <=135K  8  1  4
  > 135K  2  9  2
```

The *research hypothesis* can be made about a model: for example, since it looks like there's an association between lower salary bracket and gender, maybe the data is enough evidence to refute: H_0: gender and salary bracket are indepenent.

Testing model fit?

To test for significance, we just need to pass our 2×3 table into the appropriate function in some statistical software.

Tests of independence for $j \times k$ tables

```r
salaries <- as.table(rbind(c(8, 1, 4),
                           c(2, 9, 2)))
dimnames(salaries) <- list(
  range = c("<=135K","> 135K"),
  gender = c(" M"," F","Nb")
  )

salaries # display the data
```

```
        gender
range      M  F Nb
  <=135K   8  1  4
  > 135K   2  9  2
```

We may be interested in asking whether all genders have similar salary allocations.

The answer will be easier to *guess* at if we convert the rows to show proportions instead of counts. Here's one way of doing this.

```
salaries.prop <- prop.table(salaries, 1)
salaries.prop
```

```
        gender
range             M          F         Nb
  <=135K 0.61538462 0.07692308 0.30769231
  > 135K 0.15384615 0.69230769 0.15384615
```

By *looking* at the table we see that Female *seem to be* more likely to be in the higher salary range.

We still want to know if this difference is significant. To assess this we can use the chi-squared test (on the counts table, not the proportions table!).

```
chisq.test(salaries)
```

```
Warning in chisq.test(salaries): Chi-squared approximation may
be incorrect


    Pearson's Chi-squared test

data:  salaries
X-squared = 10.667, df = 2, p-value = 0.004828
```

We still want to know if this difference is significant. To assess this we can use the chi-squared test (on the counts table, not the proportions table!).

```r
chisq.test(salaries)
```

```
Warning in chisq.test(salaries): Chi-squared approximation may
be incorrect


    Pearson's Chi-squared test

data:  salaries
X-squared = 10.667, df = 2, p-value = 0.004828
```

```r
fisher.test(salaries,alternative="greater")
```

```


    Fisher's Exact Test for Count Data

data:  salaries
p-value = 0.004
alternative hypothesis: greater
```

Questions:

- What are the chi-squared (and Fisher's) tests doing?
- How much does this generalize to other models?

In order to answer these questions, let us compute by hand the bits and pieces that are necessary to undrestand the mechanics of these tests.

```
       gender
range            M          F         Nb
  <=135K 0.61538462 0.07692308 0.30769231
  > 135K 0.15384615 0.69230769 0.15384615
```

Observed probability table p_{obs}.

```
       gender
range    M  F Nb
  <=135K  8  1  4
  > 135K  2  9  2
```

Table of counts u.

- **Question**: What is the expected probability table? And how about the expected u?

```
       gender
range            M          F         Nb
  <=135K 0.61538462 0.07692308 0.30769231
  > 135K 0.15384615 0.69230769 0.15384615
```

Observed probability table p_{obs}.

```
       gender
range    M  F Nb
  <=135K 8  1  4
  > 135K 2  9  2
```

Table of counts u.

- **Question**: What is the expected probability table? And how about the expected u?

  ...under the model of gender $\perp\!\!\!\perp$ party affiliation?

```
      gender
range            M          F         Nb
  <=135K 0.61538462 0.07692308 0.30769231
  > 135K 0.15384615 0.69230769 0.15384615
```

Observed probability table p_{obs}.

```
      gender
range    M  F Nb
  <=135K 8  1  4
  > 135K 2  9  2
```

Table of counts u.

- **Question**: What is the expected probability table? And how about the expected u?

  . . . under the model of gender $\perp\!\!\!\perp$ party affiliation?

- In other words: compute the MLE of p_{ij} for the parameter table p, and then the table of counts u.

This is in-class work now.

```
        gender
range            M          F         Nb
  <=135K 0.61538462 0.07692308 0.30769231
  > 135K 0.15384615 0.69230769 0.15384615
```

Observed probability table p_{obs}.

```
        gender
range     M  F Nb
  <=135K  8  1  4
  > 135K  2  9  2
```

Table of counts u.

- **Question**: What is the expected probability table? And how about the expected u?

  ...under the model of gender $\perp\!\!\!\perp$ party affiliation?

- In other words: compute the MLE of p_{ij} for the parameter table p, and then the table of counts u.

This is in-class work now.

Answer: the expected u under the model of independence is:

```
        gender
range     M  F Nb
  <=135K  5  5  3
  > 135K  5  5  3
```

- How “far away” is the observed table u from the expected table $\hat{u}$?
- Whenever we say the word *expected* we mean under whatever the model is we are considering; in this case, independence of rows and columns.

Pearson’s chi-squared *goodness-of-fit* statistic

$$X^2(u) = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(u_{ij} - \hat{u}_{ij})^2}{\hat{u}_{ij}}.$$

- **Question**: What is the value of this statistic for the observed table?

```
         gender
range     M  F Nb
  <=135K  5  5  3
  > 135K  5  5  3
```

Expected table $\hat{u}$.

```
         gender
range     M  F Nb
  <=135K  8  1  4
  > 135K  2  9  2
```

Observed table u.

This is in-class work now.

- How “far away” is the observed table u from the expected table $\hat{u}$?
- Whenever we say the word *expected* we mean under whatever the model is we are considering; in this case, independence of rows and columns.

Pearson's chi-squared *goodness-of-fit* statistic

$$X^2(u) = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(u_{ij}-\hat{u}_{ij})^2}{\hat{u}_{ij}}.$$

- **Question**: What is the value of this statistic for the observed table?

```
        gender
range     M  F Nb
  <=135K  5  5  3
  > 135K  5  5  3
```

Expected table $\hat{u}$.

```
        gender
range     M  F Nb
  <=135K  8  1  4
  > 135K  2  9  2
```

Observed table u.

This is in-class work now.

Answer: $X^2 = 10.6666667$.

What are these tests computing?

Asymptotic test (chi-square)

In an **asymptotic** goodness-of-fit test, one uses the fact that X^2 has a chi-squared distribution (with known degrees of freedom) to assess whether the observed X^2 value is extreme or not.

Exact test (Fisher)

In an **exact** goodness-of-fit test, one uses the exact distribution of the statistic. . .

What are these tests computing?

Asymptotic test (chi-square)

In an **asymptotic** goodness-of-fit test, one uses the fact that X^2 has a chi-squared distribution (with known degrees of freedom) to assess whether the observed X^2 value is extreme or not.

Exact test (Fisher)

In an **exact** goodness-of-fit test, one uses the exact distribution of the statistic. . . . . . which is **what**?

```
         gender
range      M  F Nb
  <=135K   8  1  4
  > 135K   2  9  2
```

```
         gender
range      M  F Nb
  <=135K   9  0  4
  > 135K   1 10  2
         gender
range      M  F Nb
  <=135K   9  1  3
  > 135K   1  9  3
```

Conclusion? Evidence in the data? Significance?

Definition [p-value]

Refer to Chapter 5. Discuss in lecture / board.

- Read the beginning of Chapter 9. Section 9.1: Conditional inference.
  - We are *conditioning* on the row and column sums of the table.
  - These are sufficient statistics for the independence model.
  - This is a *general strategy*. . .

Summary of hypothesis testing

3 elements: hypotheses, test statistic, (rejection) decision rule .

- In practice, check assumptions to know which test to use (for example, which distribution to reference for a decision rule).
- For model/data fit, the null hypothesis is $H_0 : \mathcal{M}$ fits the data:

Goodness of fit for a model $\mathcal{M}$

we are saying that there exists a parameter value θ_{true} such that $p_{\theta_{true}} \in \mathcal{M}$ that produced the observed data. We often take $\hat{\theta}_{MLE}$ as the best guess for the true value θ_{true} under the assumption that the model is correct.

"Is the model correct? does it fit my data?" is a question you should ask *before* using the model for analytics and drawing any conclusions.

- Sometimes model fit tests are done heuristically;
- Sometimes we have formal testing procedures.
- **Next**: we will learn how computational algebra (yes, polynomial ideals!) can be used for exact tests for models for large, sparse table data!

R Resources

```
chisq.test(salaries)
```

```
Warning in chisq.test(salaries): Chi-squared approximation may be
incorrect


    Pearson's Chi-squared test

data:  salaries
X-squared = 10.667, df = 2, p-value = 0.004828
```

```
## There is an option to "simulate p value" using Monte Carlo!
chisq.test(salaries,simulate.p.value = TRUE)
```

```


    Pearson's Chi-squared test with simulated p-value (based
    on 2000 replicates)

data:  salaries
X-squared = 10.667, df = NA, p-value = 0.006997
```

Extra reading: A smaller example, 2×2

```
                  mother.smokes
birthwt.below.2500 no yes
               no  86  44
               yes 29  30
```

This is a 2×2 contingency table, a count u of cross-classified data.

The *research hypothesis* can be made about a model: for example, since it looks like there's a positive association between low birthweight and smoking status, maybe the data is enough evidence to refute:

Extra reading: A smaller example, 2×2

```
                  mother.smokes
birthwt.below.2500 no yes
               no  86  44
               yes 29  30
```

This is a 2×2 contingency table, a count u of cross-classified data.

The *research hypothesis* can be made about a model: for example, since it looks like there's a positive association between low birthweight and smoking status, maybe the data is enough evidence to refute: H_0: weight and smoking are independent.

Testing model fit?

To test for significance, we just need to pass our 2×2 table into the appropriate function in some statistical software.

Here's an example in R:

```
birthwt.fisher.test <- fisher.test(weight.smoke.tbl)
birthwt.fisher.test
```

```

    Fisher's Exact Test for Count Data

data:  weight.smoke.tbl
p-value = 0.03618
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.028780 3.964904
sample estimates:
odds ratio
  2.014137
```

Another way:

```
chisq.test(weight.smoke.tbl)
```

```

	Pearson's Chi-squared test with Yates' continuity correcti

data:  weight.smoke.tbl
X-squared = 4.2359, df = 1, p-value = 0.03958
```

- You get essentially the same answer by running the chi-squared test, but the output isn't as useful as Fisher's test.
    - In particular, you're not getting an estimate or confidence interval for the odds ratio.
    - This is why `fisher.test()` is preferred for testing 2 x 2 tables.

Resources & License

- Quick summary notes about p-values that I wrote for Stat 514.
- Read about hypothesis tests for context of the model fitting tests in these lecture notes.
- This lesson from Penn State online offers a one-page summary of Fisher's exact test for 2×2 tables, as it was developed by Sir Fisher!
- Believe it or not, there is a great 2×2 example on Wikipedia, a page which actually contains a really good explanation for this one example.

This document is created for Math/Stat 561, Spring 2023.

All materials posted on this page are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.