Lecture 8: Frequentist hypothesis testing, and contingency tables

31 October 2007

In this lecture we'll learn the following:

1. what frequentist hypothesis testing is, and how to do it;
2. what contingency tables are and how to analyze them;
3. elementary frequentist hypothesis testing for count data, including the Chi-squared, likelihood-ratio, and Fisher's exact tests;

1 Introduction to frequentist hypothesis testing

In most of science, including areas such as psycholinguistics and phonetics, statistical inference is most often seen in the form of HYPOTHESIS TESTING within the NEYMAN-PEARSON PARADIGM. This paradigm involves formulating two hypotheses, the NULL HYPOTHESIS H_0 and the ALTERNATIVE HYPOTHESIS H_A (sometimes H_1). In general, there is an asymmetry such that H_A is more general than H_0. For example, let us take the coin-flipping example yet again. Let the null hypothesis be that the coin is fair:

$$H_0 : \pi = 0.5$$

The natural alternative hypothesis is simply that the coin may have any weighting:

$$H_A : 0 \leq \pi \leq 1$$

We then design a *decision procedure* on the basis of which we either ACCEPT or REJECT H_0 on the basis of some *experiment* we conduct. (Rejection of H_0 entails acceptance of H_A.) Now, within the Neyman-Pearson paradigm the true state of the world is that H_0 is either true or false. So the combination of the true state of the world with our decisions gives the following logically possible outcomes of an experiment:

(1)

		Null hypothesis	
		Accepted	Rejected
Null Hypothesis	True	Correct decision $(1-\alpha)$	Type I error (α)
	False	Type II error (β)	Correct decision $(1-\beta)$

As you can see in (1), there are two sets of circumstances under which we have done well:

1. The null hypothesis is true, and we accept it (upper left).
2. The null hypothesis is false, and we reject it (lower right).

This leaves us with two sets of circumstances under which we have made an error:

1. The null hypothesis is true, but we reject it. This by convention is called a TYPE I ERROR.
2. The null hypothesis is false, but we accept it. This by convention is called a TYPE II ERROR.

Let's be a bit more precise as to how hypothesis testing is done within the Neyman-Pearson paradigm. We know in advance that our experiment will result in the collection of some data $\vec{x}$. Before conducting the experiment, we decide on some TEST STATISTIC T that we will compute from $\vec{x}$.[1] We can think of T as a random variable, and the null hypothesis allows us to compute the distribution of T. Before conducting the experiment, we partition the range of T into an ACCEPTANCE REGION and a REJECTION REGION.[2]

[1] Formally T is a function of $\vec{x}$ so we should designate it as $T(\vec{x})$, but for brevity we will just write T.

[2] For an unapologetic Bayesian's attitude about the Neyman-Pearson paradigm, read Section 37.1 of **?**.

(2) **Example:** a doctor wishes to evaluate whether a patient is diabetic. [Unbeknownst to all, the patient actually is diabetic.] To do this, she will draw a blood sample, $\vec{x}$, and compute the glucose level in the blood, T. She follows standard practice and designates the acceptance region as $T \leq 125\texttt{mg/dL}$, and the rejection region as $T > 125\texttt{mg/dL}$. The patient's sample reads as having $114\texttt{mg/dL}$, so she diagnoses the patient as not having diabetes, committing a Type II error.

In this type of scenario, a Type I error is often called a FALSE POSITIVE, and a Type II error is often called a FALSE NEGATIVE.

The probability of Type I error is often denoted α and is referred to as the SIGNIFICANCE LEVEL of the hypothesis test. The probabilty of Type II error is often denoted β, and $1 - \beta$, which is the probability of correctly rejecting a false null hypothesis, is called the POWER of the hypothesis test. To calculate β and thus the power, however, we need to know the true model.

Now we'll move on to another example of hypothesis testing in which we actually deploy some probability theory.

1.1 Hypothesis testing: a weighted coin

You decide to investigate whether a coin is fair or not by flipping it 16 times. As the test statistic T you simply choose the number of successes in 16 coin flips. Therefore the distribution of T under the null hypothesis H_0 is simply the distribution on the number of successes r for a binomial distribution with parameters $16, 0.5$, given below:

(3)

T	0	1	2	3	4	5	6	7	8
$p(T)$	0.0000153	0.000244	0.00183	0.00850	0.0278	0.0667	0.122	0.175	0.196
T	9	10	11	12	13	14	15	16	
$p(T)$	0.175	0.122	0.0667	0.0278	0.00854	0.00183	0.000244	0.0000153	

We need to start by partitioning the possible values of T into acceptance and rejection regions. The significance level α of the test will simply be the probability of landing in the rejection region under the distribution of T given in (3) above. Let us suppose that we want to achieve a significance level at least as good as $\alpha = 0.05$. This means that we need to choose as the rejection region a subset of the range of T with total probability mass no greater than 0.05. **Which values of T go into the rejection region is a matter of convention and common sense.**

Intuitively, it makes sense that if there are very few successes in 16 flips, then we should reject H_0. So we decide straight away that the values $T \leq 3$ will be in the rejection region. This comprises a probability mass of about 0.01:

```
> sum(dbinom(0:3,16,0.5))
[1] 0.01063538
```

We have probability mass of just under 0.04 to work with. Our next step now comes to depend on the alternative hypothesis we're interested in testing. If we are sure that the coin is not weighted towards heads but we think it may be weighted towards tails, then there is no point in putting high values of T into the rejection region, but we can still afford to add $T = 4$. We can't add $T = 5$, though, as this would put us above the $\alpha = 0.05$ threshold:

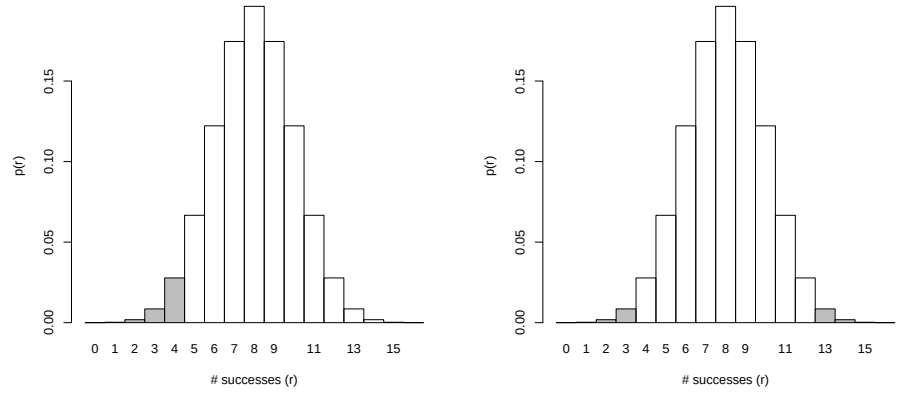
```
> sum(dbinom(0:4,16,0.5))
[1] 0.03840637
> sum(dbinom(0:5,16,0.5))
[1] 0.1050568
```

Our rejection region is thus $T \leq 4$ and our acceptance region is $T \geq 5$. This is called a ONE-TAILED TEST and is associated with the alternative hypothesis $H_A : \pi < 0.5$. We can visualize the acceptance and rejection regions as follows:

```
x <- 0:16
colors <- c(rep(8,5),rep(0,12)) # color 8 is gray, 0 is white
barplot(dbinom(x,16,0.5),names.arg=x,space=0,col=colors,
  xlab="# successes (r)", ylab="p(r)")
  # barplot() allows more flexibility in coloring
```

Alternatively, we may have no reason to believe that the coin, if unfair, is weighted in a particular direction. In this case, symmetry demands that for every low value of T we include in the rejection region, we should include a corresponding high value of equal probability. So we would add the region $T \geq 13$ to our rejection region:

```
> sum(dbinom(0:3,16,0.5),dbinom(13:16,16,0.5))
[1] 0.02127075
```



(a) One-tailed test (b) Two-tailed test

Figure 1: Acceptance and rejection regions for one- and two-tailed tests for null hypothesis that a coin is fair. Acceptance regions are white; rejection regions are gray

We cannot add 4 and 12 to our rejection region because we would wind up with $\alpha > 0.05$:

```
> sum(dbinom(0:4,16,0.5),dbinom(12:16,16,0.5))
[1] 0.07681274
```

so we are finished and have the acceptance region $5 \leq T \leq 12$, with other values of T falling in the rejection region. This type of symmetric rejection region is called a TWO-TAILED TEST, which is associated with the alternative hypothesis $H_A : \pi \neq 0.5$.[3] We can visualize this as follows:

```
x <- 0:16
colors <- c(rep(8,4),rep(0,9), rep(8,4)) # color 8 is gray, 0 is white
barplot(dbinom(x,16,0.5),names.arg=x,space=0,col=colors,
  xlab="# successes (r)", ylab="p(r)")
  # barplot() allows more flexibility in coloring
```

[3]It is actually not just convention that associates these different alternative hypotheses with one- and two-tailed tests, but also the idea that the rejection region should be chosen so as to maximize the power of the test for a pre-specified choice of α. It turns out that these extreme-value choices of the rejection region accomplish this task, but explaining how is a more detailed discussion than we have time for.

In quantitative linguistics, you will nearly always see two-tailed tests rather than one-tailed tests, because the use of one-tailed test opens the door to post-hoc "explanations" of why, *a priori*, we don't expect to see deviations from the null hypothesis in the direction that we didn't see.

1.2 One-sample t-test

Finally, we cover the one-sample t-test. Suppose we believe we are sampling normally-distributed data and we want to test the null hypothesis that the mean of the data is a certain prespecified value μ_0. We can use the fact that the "standardized" mean of the distribution is t-distributed, as in Equation (**??**), to test $H_0 : \mu = \mu_0$. We can replicate an example given by Baayen for the distribution of duration for the Dutch prefix *ont-*. `t.test()`

```
t.test(durationsOnt$DurationPrefixNasal, mu = 0.053)
[...]
t = -1.5038, df = 101, p-value = 0.1358
```

The (two-tailed) p-value for the t-statistic is 0.1358, so that we cannot reject H_0 at the $\alpha = 0.05$ level (or even the "marginal" $\alpha = 0.1$ level).

When you compare the means between two samples, the difference is also t-distributed but in a more complicated way. `t.test()` can also these two-sample tests.

1.3 Hypothesis testing: summary

- The Neyman-Pearson paradigm involves the formulation of two competing hypotheses: the null hypothesis H_0 and a more general alternative hypothesis H_A;
- H_0 and H_A are compared by choosing, in advance, a test statistic T to be calculated on the basis of data from an experiment, and partitioning the range of T into acceptance and rejection regions for H_0;
- Incorrectly rejecting H_0 when it is true is a Type I error (false positive); incorrectly accepting H_0 when it is false is a Type II error (false negative);
- The probability α of Type I error is the significance level of the hypothesis test;

- If we denote the probability of Type II error as β, then $1-\beta$ (the probability of correctly rejecting a false H_0) is the power of the hypothesis test.

2 Contingency tables

There are many situations in quantitative linguistic analysis where you will be interested in the possibility of association between two categorical variables. In this case, you will often want to represent your data as a contingency table. Here's an example from my own research, on parallelism in noun-phrase coordination, [[NP1] and [NP2]]. Consider the following four noun phrases:

> The girl and the boy (parallel; no PPs)
> The girl from Quebec and the boy (not parallel)
> The girl and [the boy from Ottawa] (not parallel)
> The girl from Quebec and the boy from Ottawa (parallel; both with PPs)

I was interested in whether NP1 and NP2 tended to be similar to each other. As one instance of this, I looked at the patterns of PP modification in the Brown and Switchboard corpora, and came up with contingency tables like this:

(4)

			NP2		
Brown			hasPP	noPP	
NP1	hasPP		95	52	147
	noPP		174	946	1120
			269	998	1267

			NP2		
Switchboard			hasPP	noPP	
NP1	hasPP		78	76	154
	noPP		325	1230	1555
			403	1306	1709

From the table you can see that in both corpora, NP1 is more likely to have a PP postmodifier when NP2 has one, and NP2 is more likely to have a PP postmodifier when NP1 has one. But we would like to go beyond that and (a) *quantify* the predictive power of knowing NP1 on NP2; and (b) *test for significance* of the association.

3 Quantifying association: odds ratios

Given a contingency table of the form

$$\begin{array}{cccc} & & Y & \\ & & y_1 & y_2 \\ X & x_1 & n_{11} & n_{12} \\ & x_2 & n_{21} & n_{22} \end{array}$$

one of the things that's useful to talk about is how the value of one variable affects the distribution of the other. For example, the overall distribution of Y is

$$freq(y_1) = \frac{n_{11}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} \quad freq(y_2) = \frac{n_{12}+n_{22}}{n_{11}+n_{12}+n_{21}+n_{22}}$$

Alternatively we can speak of the overall *odds* ω^Y of y_1 versus y_2:

$$\omega^Y \equiv \frac{freq(y_1)}{freq(y_2)} = \frac{\frac{n_{11}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}}}{\frac{n_{12}+n_{22}}{n_{11}+n_{12}+n_{21}+n_{22}}} = \frac{n_{11}+n_{21}}{n_{12}+n_{22}}$$

If $X = x_1$, then the odds for Y are just $\omega_1^Y = \frac{n_{11}}{n_{12}}$. If the odds of Y for $X = x_2$ are greater than the odds of Y for $X = x_1$, then the outcome of $X = x_2$ **increases** the chances of $Y = y_1$. We can express the effect of the outcome of X on the odds of Y by the **odds ratio** (which turns out to be symmetric between X, Y):

$$\mathcal{OR} = \frac{\omega_1}{\omega_2} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

An odds ratio $\mathcal{OR} = 1$ indicates no association between the variables. For the Brown and Switchboard parallelism examples:

$$\mathcal{OR}_{Brown} = \frac{95\times946}{52\times174} = 9.93 \quad \mathcal{OR}_{Swbd} = \frac{78\times1230}{325\times76} = 3.88$$

So the presence of PPs in left and right conjunct NPs seems more strongly interconnected for the Brown (written) corpus than for the Switchboard (spoken).

4 Testing the significance of association

In frequentist statistics there are several ways to test the significance of the association between variables in a two-way contingency table. Although you may not be used to thinking about these tests as the comparison of two hypotheses in form of statistical models, they are!

4.1 Fisher's exact test

This test applies to a 2-by-2 contingency table:

		Y		
		y_1	y_2	
X	x_1	n_{11}	n_{12}	n_{1*}
	x_2	n_{21}	n_{22}	n_{2*}
		n_{*1}	n_{*2}	n

H_0 is the model that all ***marginal totals*** are fixed, but that the individual cell totals are not – alternatively stated, that the individual outcomes of X and Y are independent. **This means that under H_0, the true odds ratio $\mathcal{OR}$is 1.** [H_0 has one free parameter – why?] H_A is the model that the individual outcomes of X and Y are not independent. With Fisher's exact test, you directly calculate the *exact* likelihood of obtaining a result as extreme or more extreme than the result that you got. [Since it is an *exact* test, you can use Fisher's exact test regardless of expected and actual cell counts.]

In R, you use the `fisher.test()` function to execute Fisher's exact test: fisher.test(),matrix()

```
> brown.nps <- matrix(c(95,174,52,946),2,2)
> fisher.test(brown.nps)

Fisher's Exact Test for Count Data

data:  brown.nps
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  6.718106 14.737945
```

Notice how R told us that H_A is the hypothesis that the true odds ratio is not equal to 1.

4.2 Chi-squared test

This is probably the contingency-table test you have heard most about. It can be applied to arbitrary two-way $m \times n$ tables, if you have a model with

k parameters that predicts expected values E_{ij} for all cells. You calculate Pearson's X^2 statistic:

$$X^2 = \sum_{ij} \frac{[n_{ij} - E_{ij}]^2}{E_{ij}}$$

In the chi-squared test, H_A is the model that each cell in your table has its own parameter p_i in one big multinomial distribution. When the expected counts in each cell are large enough (the generally agreed lower bound is ≥ 5), the X^2 statistic is distributed as χ^2_{n-k-1}.

The most common way of using Pearson's chi-squared test is to test for the independence of two factors in a two-way contingency table. Take a $k \times l$ two-way table of the form:

	y_1	y_2	$\cdots$	y_l	
x_1	n_{11}	n_{12}	$\cdots$	n_{1l}	n_{1*}
x_2	n_{21}	n_{22}	$\cdots$	n_{2l}	n_{2*}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_l	n_{k1}	n_{k2}	$\cdots$	n_{kl}	n_{k*}
	n_{*1}	n_{*2}	$\cdots$	n_{*l}	n

Our null hypothesis is that the x_i and y_i are independently distributed from one another. By the definition of probabilistic independence, that means that H_0 is:

$$P(x_i, y_j) = P(x_i)P(y_j)$$

In the chi-squared test we use the maximum-likelihood estimates $P(x_i) = P_{MLE}(x_i) = \frac{n_{i*}}{n}$ and $P(y_j) = P_{MLE}(y_j) = \frac{n_{*j}}{n}$. This gives us the formula for the expected counts:

$$E_{ij} = nP(x_i)P(y_j)$$

For the Brown corpus data in (4), we have

$$P(x_1) = \frac{147}{1267} \quad = 0.1160 P(y_1) \quad = \frac{269}{1267} = 0.2123 \tag{1}$$

$$P(x_2) \quad = 0.8840 P(y_2) \quad = 0.7877 \tag{2}$$

$$E_{11} = 31.2 \tag{3}$$

$$E_{12} = 115.8 \tag{4}$$

$$E_{21} = 237.8 \tag{5}$$

$$E_{21} = 882.2 \tag{6}$$

$$\tag{7}$$

Comparing with (4), we get

$$X^2 = (95 - 31.2)^2/31.2 + (52 - 115.8)^2/115.8 + (174 - 237.8)^2/237.8 + (946 - 882.2)^2/882.2 \tag{8}$$

$$= 187.3445 \tag{9}$$

We had 2 parameters in our model of independence, and there are 4 cells, so X^2 is distributed as χ^2_1 (4-2-1 = 1).

We can compare this with the built-in `chisq.test()` function:[4] `chisq.test()`

```
> chisq.test(matrix(c(95,174,52,946),2,2),correct=F)

Pearson's Chi-squared test

data:  matrix(c(95, 174, 52, 946), 2, 2)
X-squared = 187.2482, df = 1, p-value < 2.2e-16
```

The exact numerical result is off because I rounded things off aggressively, but you can see where the result comes from.

4.3 Likelihood ratio test

With this test, the statistic you calculate for your data D is the ***likelihood ratio***

[4] There is something called a "continuity correction" in the chi-squared test which we don't need to get into which is usable for 2×2 tables. The exposition we're giving here ignores this correction.

$$\Lambda^* = \frac{\max P(D; H_0)}{\max P(D; H_A)}$$

that is: the ratio of the maximum data likelihood under H_0 to the maximum data likelihood under H_A. This requires that you explicitly formulate H_0 and H_A. $-2 \log \Lambda^*$ is distributed like a chi-squared with ***degrees of freedom*** equal to the difference in the the number of free parameters in H_A and H_0. [Danger: don't apply this test when expected cell counts are low, like < 5.]

The likelihood-ratio test gives similar results as the chi-squared for contingency tables, but is more flexible because it allows the comparison of arbitrary nested models.