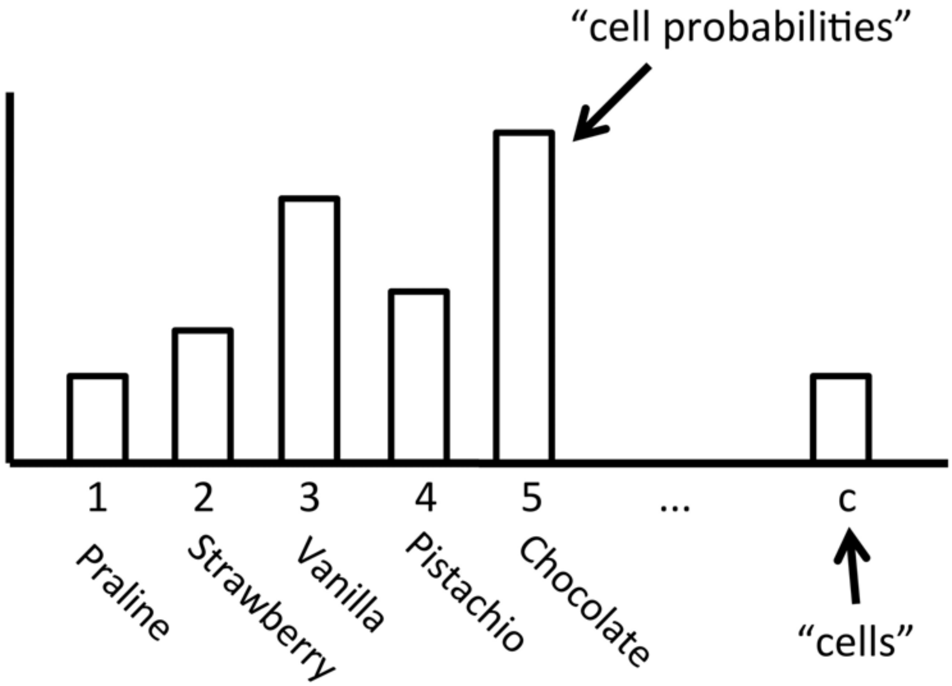
Chapter 9 Notes, 9.3 First Part
Inference for One Way Count Data
Chi-Square Test using the Multinomial Distribution

An example of the multinomial distribution: preference of ice cream flavors:



- Cells are numbered $1, \ldots, c$
- Cell probabilities are $p_1, \ldots, p_c$ where $\sum_i p_i = 1$
- Cell counts are $n_1, \ldots, n_c$ where $\sum_i n_i = n$
- Count r.v.'s $N_1, \ldots, N_c$ where $\sum_i N_i = n$.
- Multinomial distribution

$$P(N_1 = n_1, N_2 = n_2, \ldots) = \frac{n!}{n_1! n_2! \cdots n_c!} p_1^{n_1} p_2^{n_2} \cdots p^{n_c}.$$

We want to test:

$$\begin{aligned} H_0 &: p_1 = p_{10},\ p_2 = p_{20}\ , \cdots\ ,\ p_c = p_{c0} \\ H_1 &: \text{at least one } p_i \neq p_{i0} \end{aligned}$$

Example: a market survey of detergents

p_{i0} are past market shares

p_i are current market shares

$n_1, n_2, \cdots n_c$ are cell counts in the sample of the current market.

Want to test whether current shares are different from the past.

Construct test statistic χ^2 as follows:

$$e_i = np_{i0} \leftarrow \text{ expected cell counts when } H_0 \text{ is true.}$$

$$\chi^2 = \sum_{i=1}^{c} \frac{(n_i - e_i)^2}{e_i} = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

Think of χ^2 as a discrepancy of how different the observed counts are from the expected counts.

So you want χ^2 to be small. If it's too large, it means that the observed are different from the expected. If that happens, it means something has gone wrong, namely your assumption that H_0 is true. This means we'll reject H_0 if χ^2 is too large.

It is possible to show that as $n \to \infty$, χ^2 has a chi-square distribution with d.f. $c - 1$. (Note: We lost a d.f. since $\quad p_i = 1$.) So, H_0 can be rejected at level α if $\chi^2 > \chi^2_{c-1,\alpha}$.

Example: Mendel's genetic experiments

The χ^2 we introduced is a *Pearson chi-square* statistic:

$$\chi^2 = \sum \frac{(n_i - e_i)^2}{e_i} = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}.$$

Remember, this only approximately has a chi-square distribution when n is large:

$$e_i \geq 1 \text{ and more than } 4/5^{\text{th}}\text{s of } e_i\text{'s are } \geq 5. \leftarrow \text{Important}$$

15.075J / ESD.07J Statistical Thinking and Data Analysis
Fall 2011