χ^2- goodness of fit test

Some questions:

- Are the dice we roll in our experiments in class really fair ?
- Is getting Dengue(D) or severe form of Dengue (DSS) independent of `BICARB1` reading ?

Rephrase:

- How well the distribution of the data fit the model ?
- Does one variable affect the distribution of the other ?

χ^2- goodness of fit test

Specific Question:

- To understand how "close" are the observed values to those which would be expected under the fitted model ?

Towards Answer:

- In this case we seek to determine whether the distribution of results in a sample could plausibly have come from a distribution specified by a null hypothesis.
- The test statistic is calculated by comparing the observed count of data points within specified categories relative to the expected number of results in those categories (under Null).

χ^2- goodness of fit test

- Let T be a random variable with finite range $\{c_1, c_2, \ldots, c_k\}$ for which

$$P(T = c_j) = p_j > 0 \text{ for } 1 \leq j \leq k.$$

- Let $X_1, X_2, \ldots, X_n$ be the sample from the distribution T and let

$$Y_j = |\{j : X_j = c_j\}| \text{ for } 1 \leq j \leq k..$$

Y_j is the number of sample points whose outcome was c_j

- Then the statistic

$$\mathbf{X}^2 := \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \equiv \sum_{j=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

χ^2- goodness of fit test

$$\mathbf{X}^2 := \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \equiv \sum_{j=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- $\mathbf{X}^2$- has χ^2_{k-1} degrees of freedom, assymptotically as $n \to \infty$.
- Null Hypothesis: Distribution comes from Multinomial with parameters $p_1, p_2, \ldots, p_k$
- Alternate Hypothesis: Distribution comes from Multinomial with parameters with at least one parameter different from $p_1, p_2, \ldots, p_k$

χ^2- goodness of fit test

Example:

We divide the political parties in India into 3 large alliances: NDA, UPA, and Third-Front. In the previous election the support had been 38%, 32% and 30% support respectively. Super-Nation TV channel takes a sample of 100 people and finds that there are 35 for NDA, 40 for UPA and 25 for Third-Front. It concludes that the vote share has not changed. Is this hypothesis correct ?

χ^2- goodness of fit test

- Null Hypothesis: Vote Share is $(38, 32, 30)$
- Level of Significance: 0.05
- Data: Sample Vote share is $(35, 40, 25)$

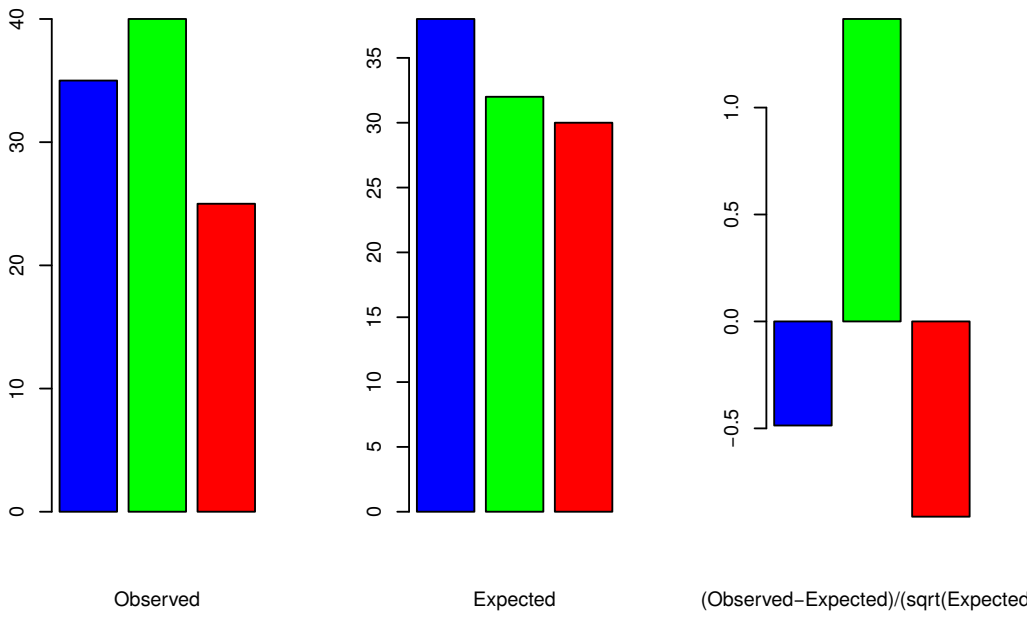
χ^2- goodness of fit test

Example Contd.:

```
> x = c(35,40,25)
> prob = c(38,32,30)
> prob = prob/sum(prob)
> n = sum(x)
> z = (x-n*prob)/((sqrt(n*prob)))
```

χ^2- goodness of fit test

Example Contd.:



χ^2- goodness of fit test

Example Contd.:

```
> Xsquared = sum(((x-n*prob)^2)/(n*prob))
> Xsquared
[1] 3.070175
> pchisq(Xsquared, df = 3 -1, lower.tail=FALSE)
[1] 0.2154368
```

Since p-value is not smaller than 0.05 we do not reject the null hypothesis.

χ^2- goodness of fit test

Example Contd.: We can use in built R function

```
> chisq.test(x,p=prob)

        Chi-squared test for given probabilities

data:  x
X-squared = 3.0702, df = 2, p-value = 0.2154
```

χ^2- goodness of fit test

$$\mathbf{X}^2 := \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \equiv \sum_{j=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- Large values of $\mathbf{X}^2$ indicate that the observed counts don't match expected counts.
- Large values of $\mathbf{X}^2$ indicates evidence that Null is not correct.

χ^2- goodness of fit test

- Test Statistic:

$$\mathbf{X}^2 := \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \equiv \sum_{j=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- Decide on level of significance: α
- Compute p-value:

$$\mathbb{P}(\chi^2_{k-1} \geq X^2)$$

- Reject Null Hypotheis:

if p-value is less than α