Review of Statistics

January 7, 2003

Every study or Experiment yields a set of data.

Central Tendency

The center of the data set or the point about which the observations tend to cluster.

Example:

Forced Expiratory Volumes (FEV) in 1 second for asthma patients:

x_1	2.3	x_5	2.75	x_9	2.68	x_{13}	3.38
x_2	2.15	x_6	2.84	x_{10}	3.00		
x_3	3.50	x_7	4.04	x_{11}	4.02		
x_4	2.60	x_8	2.25	x_{12}	2.85		

x_i = single measurement when i can take any value from 1 to $n = 13$:

$$
\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{n}(x_1 + x_2 + \ldots + x_n) \\
&= \frac{1}{13}(2.30 + \ldots + 3.38) \\
&= \frac{38.35}{13} \\
&= 2.95 \text{ liters}
\end{aligned}
$$

Measure of Dispersion

The measure that describe the variability in a set of data values. The most commonly used measures are variance and standard deviation. The variance is:

$$S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

For FEV data the variance and standard deviation are:

$$\begin{aligned} S^2 &= \frac{1}{(13-1)} \sum_{i=1}^{13} (x_i - 2.95)^2 \\ &= \frac{1}{12}[(2.30 - 2.95)^2 + \ldots + (3.38 - 2.95)^2] \\ &= \frac{4.66}{12} \\ &= 0.39 \text{ liters}^2 \end{aligned}$$

$$\begin{aligned} S &= \sqrt{S^2} \\ &= 0.62 \text{ liters.} \end{aligned}$$

Tests of Hypothesis

Hypothesis is a claim or statement either about the value of a single population characteristic or about the values of several population characteristics.

Null hypothesis — H_0

—is the claim about a population characteristic that is initially assumed to be true.

Alternative hypothesis — H_a — is the competing claim.

Test procedure— the decision rule that is used for determine whether H_0 should be rejected.

There is some chance that the use of a test procedure for a sample data lead us to a wrong conclusion.

Type I error the error of rejecting H_0 when H_0 is true.

Type II error the error of failing to reject H_0 when H_0 is false.

Probability of type I error $= \alpha$ — called level of significance of the test.

The probability of type II error $= \beta$.

The fundamental idea behind hypothesis testing procedure is

We reject the null hypothesis if the observed sample is very unlikely to have occurred when H_0 is true.

Test Statistic—the quantity used as a basis for our decision. A test statistic is the function of sample data on which a conclusion to reject or fail to reject H_0 is based.

Example: Suppose somebody would like to test the claim that the mean Forced Expiratory volume in 1 second for asthma patient is 3.50 liters, i.e., would like to test the null hypothesis

$$H_0 : \mu = 3.50 \text{ liters}$$

The test statistic used is called t-statistic

$$\begin{aligned} t &= \frac{(\bar{x} - \mu)}{s/\sqrt{n}} \\ &= \frac{(2.95 - 3.5)}{.6/\sqrt{13}} \\ &= -3.305089 \end{aligned}$$

- Let level of significance be $\alpha = 0.05$.

- $t \sim t_{n-1}$ distribution

- $t_{.05}(n-1) = t_{.05}(12) = -1.782288$

- Since $t < t_{.05}(12)$ reject H_0.

Chi-square Statistic: – Goodness of fit tests and independence tests for discrete variables can be formulated as chi-square tests. For goodness of fit test, the test statistic is:

$$\chi^2 = \sum_{i=1}^{a} \frac{(o_i - e_i)^2}{e_i}$$

where

a : is number of classes.

o_i : is observed frequency for class i.

e_i : expected counts for class i.

Example: Tossing a die 48 times. Is die fair? The null hypothesis is: $H_0 : p(1) =, \ldots, = p(6) = 1/6$.

i	1	2	3	4	5	6
o_i	10	6	8	10	6	8
e_i	8	8	8	8	8	8

The $\chi^2 = \sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i} = 2$

$\chi^2_{1-\alpha} = \chi^2_{.95}(5) = 11.07$, $2 < \chi^2_{.95}(5)$. Do not reject H_0 and concluded that the die is fair.

p-value

The p-value is a measure of inconsistency between the hypothesized value for a particular characteristic and the observed sample.
It is the probability, assuming that H_0 is true, of obtaining a test statistic value at least as contradictory to H_0 as what usually expected.
A decision as to whether or not H_0 should be rejected results from comparing the p-value to the chosen α.
H_0 should be rejected if p-value $\leq \alpha$
H_0 should not rejected if p-value $> \alpha$.
$p - value = p(t < -3.305089) = 0.0031$ for $H_a : \mu < 3.5$. Reject H_0.
$p{-}value = p(\chi^2 > 2) = 0.85$ for H_a : die is not fair. Do not Reject H_0.