SMGroup

SM Journal of Biometrics & Biostatistics

Short communication

What is "P-value" and How to get it?

Po-Huang Chyou*

Biomedical Informatics Research Center at Marshfield Clinic Research Foundation, Marshfield, USA

Article Information

Received date: Aug 03, 2017
Accepted date: Aug 30, 2017
Published date: Sep 04, 2017

***Corresponding author**

Po-Huang Chyou, Biomedical Informatics Research Center at Marshfield Clinic Research Foundation, Marshfield, USA, Email: chyou.po-huang@marshfieldresearch.org

Let's start with an example. A random sample of 400 persons included 240 smokers and 160 non-smokers. Of the smokers, 192 had Coronary Heart Disease (CHD), while only 32 non-smokers had CHD. Could a health insurance company claim the proportion of smokers having CHD differs from the proportion of non-smokers having CHD? This is a typical hypothesis testing problem. In general, there are 6 steps for performing hypothesis testing. Step 1: define the null hypothesis (H_0). Step 2: define the alternative hypothesis (H_a). Step 3: define the type I error (α) and sample size (n). Step 4: define a statistic and the rejection region. Step 5: calculate the statistic using the sample data. Step 6: state the conclusion (reject H_0 or not). For the above example, let us assume P_1 represents the true proportion of smokers having CHD and P_2 is the true proportion of non-smokers having CHD. Then, Step 1: forming the null hypothesis H_0: $P_1 = P_2$. Step 2: forming the alternative hypothesis H_a: $P_1 \neq P_2$. Step 3: we select α = .05 and we know n = 400. Step 4: for comparing the difference in two proportions we choose statistic $z = (p_1 - p_2)/\text{sqrt}\ (p(1 - p)*(1/n_1 + 1/n_2))$, where p_1 = sample proportion of smokers having CHD = x_1/n_1 = 192/240 = .80, p_2 = sample proportion of non-smokers having CHD = x_2/n_2 = 32/160 = .20, p = overall sample proportion of total subjects (i.e., both smokers and non-smokers) having CHD = $(x_1 + x_2)/(n_1 + n_2)$ = (192 +32)/(240 + 160) = 224/400 = 0.56 and "sqrt" in the statistic z formula denotes taking the square root. Therefore, in Step 5 we calculate our statistic z = (.80 - .20)/sqrt ((.56) (1 - .56)*((1/240 + 1/160))) = .60/.05066 = 11.84. Since 11.84 exceeds the rejection region value of 1.96, in Step 6 we reject H_0 and conclude that smokers had significantly higher proportion of CHD than that of non-smokers (P-value < .0000001).

So, what is "P-value"? Here is our definition 1: a P-value is the likelihood (in probability) of incorrectly rejecting the first (H_0) hypothesis based on the data you have collected, received, or generated through simulation. Question is: Do you need a "P-value"? The answer is "No" if you don't perform any hypothesis testing, but "Yes" if you do perform some hypothesis testing. In statistics, many hypothesis testing can be considered. For instance, hypothesis testing on population mean(s), population median(s), population proportion(s), population variance(s), population correlation(s), association based on contingency table(s), coefficients based on regression model, odds ratio, relative risk, trend analysis, survival distribution(s)/curve(s), goodness of fit, just name a few.

There are 2 types of error in statistics which are relevant to what is "P-value" that are needed to be described here. First, type I error: we reject H_0 but H_0 is true. That is, α = Pr (reject H_0/H_0 is true) = Pr (type I error) = Level of significance in hypothesis testing. "Pr" is the abbreviation of "Probability" and "/" means "given". Second, type II error: we accept H_0 but H_0 is false. That is, β = Pr (accept H_0/H_0 is false) = Pr (type II error). With one of these types of error, what is "P-value" can also be defined as: Definition 2 – A P-value is the "smallest type I error" for which the first (H_0) hypothesis is rejected based on the data you have collected, received, or generated through simulation.

Notice that, in the example above, no explanations on how to come up with "P-value < .0000001" was given. The next description refers to the second part of our topic: How to get "P-value". To do that, it is informative to refresh Steps of hypothesis testing: Step 1 - Formulate the null hypothesis H_0 in statistical terms; Step 2 - Formulate the alternative hypothesis H_a in statistical terms; Step 3 - Set the level of significance α and the sample size n; Step 4 - Select the appropriate statistic and the rejection region; Step 5 - Collect the data and calculate the statistic; and Step 6 - If the calculated statistic falls in the rejection region, reject H_0 in favor of H_a; if the calculated statistic falls outside the rejection region, do not reject H_0. One key question is: What is "test statistic"? There exist a lot of distributions in statistics. For the discrete distributions the following three are most commonly encountered: 1) Binomial - $(n!/(x!*(n - x)!)\ p^x(1 - p)^{n - x}$, where x = 0, 1, 2,…, n; 2) Trinomial - $(n!/(x_1!x_2!*(n - x_1 - x_2)!p_1^{x1}p_2^{x2}*(1 -p_1 - p_2)^{n - x1 - x2})$, where x_1, x_2 = 0, 1, 2,…, n and $x_1 + x_2 < n$; 3) Poisson - $\lambda^{x}*e^{-\lambda}/x!$, where 0 < λ and x = 0, 1, 2,… (Example: 3! = 3*2*1 = 6). On the other hand, for the continuous distributions, four are as follows: 1) z - $1/(\sigma*\text{sqrt}(2\pi)*e^{(x - \mu)2/(2\sigma2)}$; 2) t - $\Gamma[(r + 1)/2]/(\text{sqrt}(\pi r)\Gamma(r/2)(1 + x^2/r)^{(r + 1)/2})$; 3) X^2 (chi-square) - $1/(\Gamma(r/2)2^{r/2})*x^{(r/2 - 1)}*e^{-x/2}$; and 4) F - $\Gamma[(r_1 + r_2)/2]*(r_1/r_2)^{r1/2}x^{r1/2-1}/(\Gamma(r_1/2)\Gamma(r_2/2)(1 + r_1x/r_2)^{(r1+r2)/2})$. Regarding Γ function, it is Γ(x) = (x - 1) (x - 2) (x - 3)…*2*1. For example, Γ (4) = 3*2*1 = 6. For the above-mentioned continuous distributions, test statistics have been theoretically derived as follows: 1) z = (xbar - μ_0)/(σ/sqrt(n)) where x bar is the sample mean; 2) t = (xbar - μ_0)/(s/sqrt(n)) where s is the sample standard deviation; 3) $X^2 = \sum(O_i - E_i)^2/E_i$, where O_i and E_i are the observed and expected frequencies respectively, for i = 1, 2,…,

k where k is the number of categories for ak*1 contingency table; and 4) $F = s_1^2/s_2^2$, the ratio of two sample variances.

Once the distribution of a statistic is determined and a test statistic is calculated through sample data, then via using commercially available statistical software package(s) which includes, but not limited to, PASS 13 (NCSS 2015) [1,2] which what the author used for all the calculated P-values shown in this paper, the derivation of P-value is straightforward. The "how to" will be shown by data generated through simulation process.

Data Generated Through Simulation

Our goal here is to demonstrate that P-value can be derived using data generated through simulation. Note that P-value can definitely be derived from real world data. For the preparation of data simulation, the following SAS code is required:

```
Data one;
 Do i=1 to 10;
  X=normal (1000);
  Output;
 End;
 Proc print;
Run;
```

By running the above SAS code, the following results are generated from the SAS system:

Obsi	x	
1	1	0.23441
2	2	-0.49978
3	3	-0.17211
4	4	-0.07346
5	5	-0.62066
6	6	1.03902
7	7	0.54514
8	8	0.94261
9	9	0.89600
10	10	-0.73138

Given these simulation data points (n = 10), we can now perform the following hypothesis testing:

Step 1 – H_0: $\mu = 0$ (μ_0)

Step 2 – H_a: $\mu \neq 0$ (μ_0)

Step 3 – $\alpha = .05$, $n = 10$

Steps 4-5 - Two possible statistics can be used here; they are statistics z and t, as described earlier. If the former was used, then statistic $z = (xbar - \mu_0)/\sigma = (.1560 - 0)/1 = .1560$, where σ is the standard error of xbar from a standardized normal distribution. From PASS 13, or standard normal z table, P-value = .5620. If latter was used, then statistic $t = (xbar - \mu_0)/(s/sqrt(n)) = (.1560 - 0)/(.6737/sqrt(10)) = .7322$. From PASS 13, or t table, P-value = .7587.

Step 6 - Do not reject H_0 and conclude that the mean is not significantly different from zero [P-value > .50].

Some Discussions

Gelman in his commentary [3] discussed what a P-value in practice is. He also cautioned about a misleading P-value. However, how to get P-value was not covered at all. In our study, this issue was particularly mentioned and illustrated using a simulation data set. We also attempted to clearly define what P-value is and what the necessary steps of hypothesis testing are for the purposes of deriving a P-value and drawing a valid conclusion based on the calculated P-value.

In summary, with clear understanding of what P-value is and what it is for, as well as with available proper statistical software package(s) in hand, and with little or no help from biostatistician(s), non-statistically trained medical professionals should be able to derived the needed P-value with ease when performing their own hypothesis testing through the application of the easy-to-adopt 6 steps as described in this paper.

References

1. PASS 13 Power Analysis and Sample Size. NCSS Statistical Software 2015 (website).
2. Hogg RV, Craig AT. Introduction to Mathematical Statistics. 4th edn. New York: Macmillan Publishing Co Inc. 1978
3. Gelman A. P Values and statistical practice. Epidemiology. 2013; 24: 69-72.