Hypothesis Testing

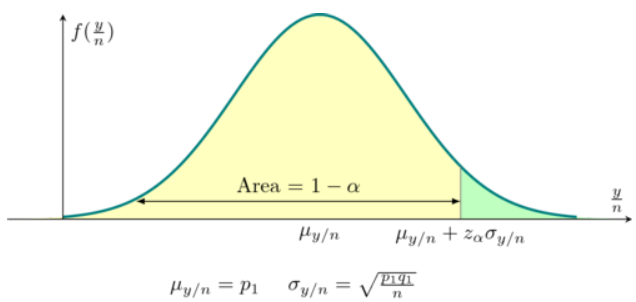
Parthiban Rajendran

October 12, 2018

As seen earlier in confidence intervals, using Wald's method for the sample proportions do not yield promising results as widely believed. So we will only stick to case when conditions are met to make the sampling distribution normalcy good enough.

1 When sample sizes are high

Suppose that we have a normal **sampling distribution** described by random variable $\dfrac{Y}{n} = N\Big(p_1, \dfrac{p_1q_1}{n}\Big)$ created from a population distribution which is a Bernoulli distribution with mean p_1 and standard deviation p_1q_1. Note that Y represents the sum of *successes* in a sample set, and thus $\dfrac{Y}{n}$ represents sample proportions. For example, for any *kth* sample set of $\dfrac{Y}{n}$, we calculate sample proportion statistic, $\dfrac{Y_k}{n} = \dfrac{1}{n}\sum\limits_{i=1}^{n} Y_{ki}$, where Y_{ki} is ith sample in kth sample set of sampling distribution described by $\dfrac{Y}{n}$. If α is the significance level, then we could derive the conditions for hypothesis testing as follows. Below is our sampling distribution as null hypothesis, with α as significance level. This is for alternate hypothesis being $H_a : \mu > \mu_{y/n}$ so we consider the right tail area. One could try similar approach for left or both tails depending on if H_a is $H_a : \mu < \mu_{y/n}$ or $H_a : \mu \neq \mu_{y/n}$ respectively.



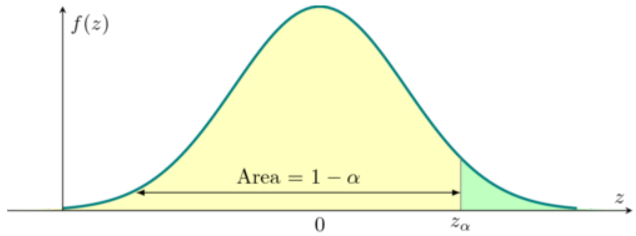
The significance level α, corresponds to the rest of $1-\alpha$ area, that is green area as shown above.

$$P\Big(\frac{Y}{n} \geq \mu_{y/n} + z_\alpha\sigma_{y/n}\Big) = \alpha$$

$$\therefore P\left(\frac{\frac{Y}{n} - \mu_{y/n}}{\sigma_{y/n}} - \geq z_\alpha\right) = \alpha$$

$$P\left(\frac{\frac{Y}{n} - p_1}{\sqrt{\frac{p_1q_1}{n}}} - \geq z_\alpha\right) = \alpha$$

Let the z score be, $z = \dfrac{\frac{Y}{n} - \mu_{y/n}}{\sigma_{y/n}}$, then $P(z \geq z_\alpha) = \alpha$



Our allowed critical region in sampling distribution is $(\mu_{y/n} + z_\alpha \sigma y/n, \infty)$, where the probability of making Type I error is α. Our allowed critical region in *standardized* sampling distribution would be (z_α, ∞). So if our z score falls within (z_α, ∞), we could reject the null hypothesis. This is also equivalent to saying, if our sample set proportion y/n falls within $(\mu_{y/n} + z_\alpha \sigma y/n, \infty)$, we could reject the null hypothesis.

Conditions

- One of the main condition to apply hypothesis testing to sample proportions is to ensure the sampling distribution is normal. This is usually ensured when $(np, nq) > 10$ if not population is already normal.
- You see, unlike sample means, there was no σ not known case in proportions, because we are testing against hypothesized mean p_1, so the associated σ would be simply $\sqrt{p_1 q_1}$. So p_1 is a pre requisite against which we need to test, so that is usually given or implicit in case of one proportion, so no σ unknown case arises here.

Example

It was claimed that many commercially manufactured dice are not fair because the "spots" are really indentations, so that, for example, the 6-side is lighter than the 1-side. To test, in an experiment, several such dice were rolled, to yield a total of $n = 8000$ *observations, out of which 6 resulted,* 1389 *times. Is there a significant evidence that dice favor a 6 far more than a fair die would? Assume* $\alpha = 0.05$

Solution:

Let us assume null hypothesis as a fair die, nothing to doubt about. The probability of getting a 6 in fair die is $p = 1/6$. So

$H_0 : \mu_{y/n} = p = 1/6$
$H_a : \mu_{y/n} = p \neq 1/6$

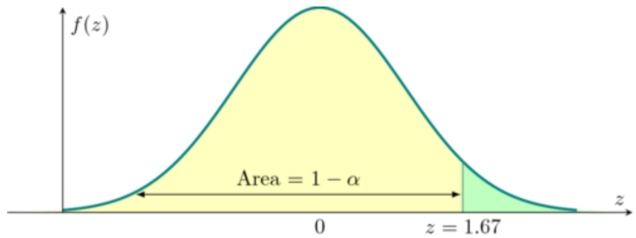
We have a sample size of $n = 8000$, so $np = (8000)(1/6) = 1333 >> 10, nq = (8000)(5/6) = 6666 >> 10$, so our normal condition is met. If we continue with sample sets of this size, we would get a good normal sampling distribution $\frac{Y}{n}$

Our z score is $z = \frac{\frac{Y}{n} - p_1}{\sqrt{\frac{p_1 q_1}{n}}} = \frac{(1389/5000) - (1/6)}{\sqrt{\frac{(1/6)(5/6)}{8000}}}$

```
In[11]: Y,n,p_1,q_1 = 1389, 8000, 1/6,5/6
        num = (Y/n) - (p_1)
        from math import sqrt
        den = sqrt(p_1*q_1/n)
        zs = round(num/den, 4)
        print(zs)
```

```
1.67
```

Our **allowed** critical region starts from $z_{0.05} = 1.645$. The z score $z = 1.67$ is greater that that, which means, if we select this sample set as critical region's starting point, our probability of making Type I error is smaller than allowed $\alpha = 0.05$. So we **reject the null hypothesis** , thus suggesting there is stronger evidence for alternate H_a.



2 Conditions Summary