I06 - Pvalues

STAT 401 (Engineering) - Iowa State University

February 21, 2018

Statistical hypothesis testing

Definition

A (classical) hypothesis test consists of two hypotheses:

- null hypothesis (H_0) and
- an alternative hypothesis (H_A)

which make a claim about parameters in a model and a decision to either

- reject the null hypothesis or
- fail to reject the null hypothesis.

We reject the null hypothesis if our pvalue is less than a pre-determined significance level a where the pvalue is the probability *when the data are considered random* of observing a test statistic as or more extreme than that observed if the null hypothesis is true.

Binomial model

If $Y \sim Bin(n,\theta)$, then the standard hypotheses are

- $H_0 : \theta = \theta_0 = 0.5$ and
- $H_A : \theta \neq \theta_0$.

In this case, the

- test statistic is Y,
- its sampling distribution *when the null hypothesis is true is* $Y \sim Bin(n,\theta_0)$, and
- the *as or more extreme* region is values farther from $n\theta_0$ than y.

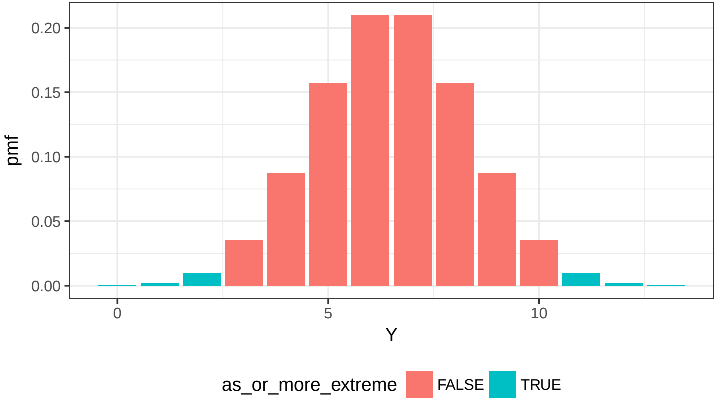
So the pvalue is

$$pvalue = P(|Y - n\theta_0| \geq |y - n\theta_0|)$$

where y is the observed successes.

```
library(dplyr); library(ggplot2)
n <- 13; y <- 2; theta0 <- 0.5
d <- data.frame(Y = 0:n) %>%
  mutate(pmf = dbinom(Y, n, theta0),
         as_or_more_extreme = abs(Y-n*theta0) >= abs(y-n*theta0))

ggplot(d, aes(Y, pmf, fill=as_or_more_extreme)) + geom_bar(stat = "identity") +
  theme_bw() + theme(legend.position="bottom")
```



Binomial example

If $Y \sim Bin(n, \theta)$ with $n = 13$ and $y = 2$ and we are testing

- $H_0 : \theta = 0.5$ versus
- $H_A : \theta \neq 0.5$,

then the pvalue is

$$pvalue = \sum_{y=0}^{2} P(Y = y|\theta = 0.5) + \sum_{11}^{13} P(Y = y|\theta = 0.5)$$

which is

```
(p <- sum(dbinom(c(0:2,11:13), size = 13, prob = 0.5)))


[1] 0.02246094
```

Thus, we would *reject the null hypothesis* for any significance level greater than 0.0224609.

binom.test

The R function ‘binom.test‘ can perform this test for us:

```
binom.test(2,13)


Exact binomial test

data:  2 and 13
number of successes = 2, number of trials = 13, p-value = 0.02246
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.01920667 0.45447106
sample estimates:
probability of success
             0.1538462
```

One-sided pvalues

If $Y \sim Bin(n, \theta)$, a one-sided hypothesis test is

- $H_0 : \theta \geq \theta_0 = 0.5$ and
- $H_A : \theta < \theta_0$.

In this case, the

- test statistic is Y,
- its sampling distribution *when the null hypothesis is true is* $Y \sim Bin(n, \theta_0)$, and
- the *as or more extreme* region is values farther from $n\theta_0$ than y in the direction of H_A.

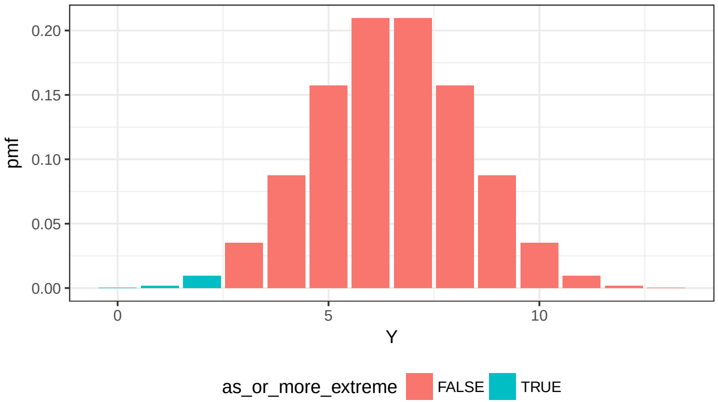
So the pvalue is

$$pvalue = P(Y - n\theta_0 \leq y - n\theta_0) = P(Y \leq y)$$

where y is the observed successes.

```
library(dplyr); library(ggplot2)
n <- 13; y <- 2; theta0 <- 0.5
d <- data.frame(Y = 0:n) %>%
  mutate(pmf = dbinom(Y, n, theta0),
         as_or_more_extreme = Y <= y)

ggplot(d, aes(Y, pmf, fill=as_or_more_extreme)) + geom_bar(stat = "identity") +
  theme_bw() + theme(legend.position="bottom")
```



Binomial example

If $Y \sim Bin(n, \theta)$ with $n = 13$ and $y = 2$ and we are testing

- $H_0 : \theta \geq 0.5$ versus
- $H_A : \theta < 0.5$,

then the pvalue is

$$pvalue = \sum_{y=0}^{2} P(Y = y|\theta = 0.5)$$

which is

```
(p <- sum(dbinom(0:2, size = 13, prob = 0.5)))

[1] 0.01123047
```

Thus, we would *reject the null hypothesis* for any significance level greater than 0.0112305.

binom.test()

The R function ‘binom.test()‘ can perform this test for us:

```
binom.test(2, 13, alternative="less")


Exact binomial test

data:  2 and 13
number of successes = 2, number of trials = 13, p-value = 0.01123
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.4100986
sample estimates:
probability of success
             0.1538462
```

Asymptotic pvalues

If we have an asymptotically normal estimator $\hat{\theta} = \hat{\theta}(Y)$, i.e.

$$\hat{\theta}(Y) \dot{\sim} N(E[\hat{\theta}], Var[\hat{\theta}])$$

then we can calculate pvalues using this approximate sampling distribution.

- $H_0 : \theta = \theta_0 \implies pvalue = P(|\hat{\theta}(Y) - E[\hat{\theta}]| \geq |\hat{\theta}(y) - E[\hat{\theta}]|)$
- $H_0 : \theta \geq \theta_0 \implies pvalue = P(\hat{\theta}(Y) \leq \hat{\theta}(y))$
- $H_0 : \theta \leq \theta_0 \implies pvalue = P(\hat{\theta}(Y) \geq \hat{\theta}(y))$

where

- $\hat{\theta}(Y)$ is the random estimator and
- $\hat{\theta}(y)$ is the observed estimator.

Binomial example

If $Y \sim Bin(n,\theta)$ and n is large (and y is not close to 0 or n), then

$$Y \dot{\sim} N(n\theta, n\theta(1-\theta)).$$

If we have

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_A : \theta \neq \theta_0,$$

then we our pvalue is

$$\begin{aligned} pvalue &= P\left(|Y - n\theta_0| \geq |y - n\theta_0|\right) \\ &= 2P\left(\frac{Y-n\theta_0}{Var[\theta]} < \frac{-|y-n\theta_0|}{SE[\hat{\theta}]}\right) \\ &\approx 2P\left(Z < \frac{-|y-n\theta_0|}{\sqrt{n\theta_0(1-\theta_0)}}\right) \end{aligned}$$

```
n = 10000; y = 4900; theta0 = 0.5
2*pnorm(-abs(y-n*theta0)/sqrt(n*theta0*(1-theta0)))
```

```
[1] 0.04550026
```

prop.test()

For the binomial distribution, the prop.test() function performs these hypothesis tests. For example, if $Y \sim Bin(n,\theta)$ and you want to test $H_0 : \theta = 0.5$ vs $H_A : \theta \neq 0.5$ when observing $y = 4900$ successes out of $n = 10^4$ attempts, the code is

```
prop.test(y, n, p = theta0, correct = FALSE)


1-sample proportions test without continuity correction

data:  y out of n, null probability theta0
X-squared = 4, df = 1, p-value = 0.0455
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4802079 0.4997998
sample estimates:
   p
0.49
```

But you should always use the continuity correction:

```
prop.test(y, n, p = theta0, correct = TRUE)$p.value


[1] 0.04659094
```

Normal mean

Let $Y_i \stackrel{ind}{\sim} N(\mu, \sigma^2)$, then

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}(0,1)$$

is our test statistic and its sampling distribution. We have the following null hypothesis tests and pvalues

- $H_0 : \mu = \mu_0$ and $pvalue = P(|T| \geq |t|) = 2P(T < -|t|)$
- $H_0 : \mu \geq \mu_0$ and $pvalue = P(T \leq t) = P(T < t)$
- $H_0 : \mu \leq \mu_0$ and $pvalue = P(T \geq t) = 1 - P(T < t)$

where

$$t = \frac{\overline{y} - \mu}{s/\sqrt{n}}$$

is the observed value of our test statistic. This is called a one-sample t-test.

t.test

```
set.seed(20180221); y <- rnorm(15, mean = 1)
t.test(y)


One Sample t-test

data:  y
t = 3.7279, df = 14, p-value = 0.002249
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.4282142 1.5884593
sample estimates:
mean of x
 1.008337
```

```
t.test(y, mu = 1, alternative = "greater")


One Sample t-test

data:  y
t = 0.38377, df = 14, p-value = 0.3535
alternative hypothesis: true mean is greater than 1
95 percent confidence interval:
 0.6380276       Inf
sample estimates:
mean of x
 1.100843
```

Relationship to confidence intervals

There is a one-to-one correspondence between pvalues and confidence intervals. Consider the following null hypotheses and corresponding confidence intervals (CIs)

- $H_0 : \theta = \theta_0$ (two-sided CI),
- $H_0 : \theta \geq \theta_0$ (one-sided lower CI), and
- $H_0 : \theta \leq \theta_0$ (one-sided upper CI),

Theorem

The appropriate (two-sided or one-sided in the correct direction) $100(1-a)\%$ confidence interval contains θ_0 if and only if the pvalue is greater than a.

Interpreting pvalues

We teach students to say the phrases

- if $pvalue < a$, reject the null hypothesis or
- if $pvalue \geq a$ fail to reject the null hypothesis.

But this is incorrect!

According to the American Statistical Association Statement on Pvalues:

> *Pvalues can indicate how incompatible the data are with a specific statistical model.*

The specific statistical model is the model associated with the null hypothesis, e.g. $Y_i \stackrel{ind}{\sim} N(\mu_0, \sigma^2)$.

So, we are not going to compare pvalues to a significance level. Instead, we are going to let pvalues mean what they meant to Sir R. A. Fisher, i.e. they indicate how incompatible the data are with a specific statistical model. (Although Fisher did suggest a cutoff of 0.05 as being "statistically significant", but he was not willing to say "reject the null hypothesis").