Workshop on GWAS

Introduction to p-values, using simulation of flipping a coin

Prof. Lei Sun

Department of Statistical Sciences, FAS

Division of Biostatistics, DLSPH

University of Toronto

23 June, 2021

Context, the scientifc question of interest

We have a coin and we want to determine if it is fair or not.

An Example

A coin is tossed 100 times and 57 heads are observed.

Is this a fair coin?

Notations

- n: the total number of tosses, a pre-specified sample size
- x: the number of heads out of n, the observed data
- theta (θ): the probability of heads, the parameter that we do not know the true value but we want to use the observed data to infer (statistical inference or learning)
- theta.0 (θ_0): what we believe theta (θ) to be (e.g. 0.5), the null hypothesis, theta=theta.0 ($\theta = \theta_0$)
- theta.hat ($\hat{\theta}$): what the data tells us (e.g. $\frac{x}{n}$, the estimate or estimator depending on the setting) about theta (θ)?

Recall the Example

A coin is tossed $n = 100$ times and $x = 57$ heads are observed.

Is this a fair coin?

The Coin Example, the sample size, observed data and null hypothesis

The sample size, n

```
n=100
```

The observed data, x

```
x=57
```

The parameter (θ) value specified by the null hypothesis, θ_0

```
theta.0=0.5
```

The Coin Example, parameter estimation (estimate, estimator)

The estimate $\hat{\theta}$
(different from an estimator if x is viewed as a random variable)

```
theta.hat=x/n
print(c(theta.hat, theta.0))
```

```
## [1] 0.57 0.50
```

theta.hat ($\hat{\theta}$) is clearly different from theta.0 (θ_0),
BUT, it seems strange if we claim the coin is not fair.

The Coin Example, summary so far

A coin is tossed $n = 100$ times and $x = 57$ heads are observed.

$$\hat{\theta} = \frac{x}{n} = \frac{57}{100} = 0.57 \longleftarrow \text{point estimate}$$

Is this a fair coin?

$$H_0 : \theta = \theta_0 = 0.5 \longleftarrow \text{hypothesis testing}$$

What do we expect $\hat{\theta}$ or x to be if the coin is fair?

To do this, we need to understand the behavior (distribtuion) of the test statistic ($\hat{\theta}$) under the null (θ_0).

We can use a R function, rbinom, that allows us to draw x from a fair coin based on n tosses.

Use `?rbinom` to understand this function (binomial distribution).

```
?rbinom
```

```
The Binomial Distribution

This is conventionally interpreted as the number of ‘successes’ in size trials.

rbinom(n, size, prob)

  n      number of observations. # number of experiments, n.rep

  size   number of trials.       # n, number of tosses in each experiment
```

Binomial Distribution, $X \sim Binom(n, \theta)$

$$\text{Prob}(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Mean (on average): $\mu = E(X) = n\theta$

Variance (how variable): $\sigma^2 = Var(X) = n\theta(1-\theta)$
Standard Deviation: $\sigma = \sqrt{n\theta(1-\theta)}$

If $n = 100$ and $\theta = 0.5$ (i.e. the coin is fair)

$$\text{Prob}(X = x) = \binom{100}{x} 0.5^x (1-0.5)^{100-x}$$

$$\mu = E(X) = 100 \cdot 0.5 = 50$$

$$\sigma^2 = Var(X) = 100 \cdot 0.5 \cdot (1-0.5) = 25$$

$$\sigma = \sqrt{Var(X)} = 5$$

If $X = x = 57$

$$\text{Prob}(X = 57) = \binom{100}{57} 0.5^{57} (1-0.5)^{43}$$

Why not just calculating the probability of the event?

$$\text{Prob}(X = 57) = \binom{100}{57} 0.5^{57}(1 - 0.5)^{43}$$

```
choose(100,57)*0.5^57*0.5^43
```

```
## [1] 0.03006864
```

Even if $x = 50$, the probability is small

$$\text{Prob}(X = 50) = \binom{100}{50} 0.5^{50}(1 - 0.5)^{50}$$

```
choose(100,50)*0.5^50*0.5^50
```

```
## [1] 0.07958924
```

Looking ahead: Probability $\neq$ Likelihood $\neq$ **p-value**

Let's do one experiment of tossing a truly fair coin n times with the probability being θ_0

This makes sure that every time we run the R program, we have the same results. It can be any number.

```
set.seed(1234)
```

```
rbinom(1,n,theta.0)
```

```
## [1] 47
```

We can run `rbinom(1,n,theta.0)` several times to check out the values we get for x

```
rbinom(1,n,theta.0)
```

```
## [1] 40
```

```
rbinom(1,n,theta.0)
```

```
## [1] 48
```

```
rbinom(1,n,theta.0)
```

```
## [1] 47
```

```
rbinom(1,n,theta.0)
```

```
## [1] 48
```

```
rbinom(1,n,theta.0)
```

```
## [1] 53
```

```
rbinom(1,n,theta.0)
```

```
## [1] 52
```

Many Experiments More Efficiently

To do this efficiently say, 10,000 times, we can use n_{rep} to specify the number of times we want to run the experiment; each time is n tosses.

```
n.rep=10000
x.simulated=rbinom(n.rep,n,theta.0)
```

Check the first 10 results

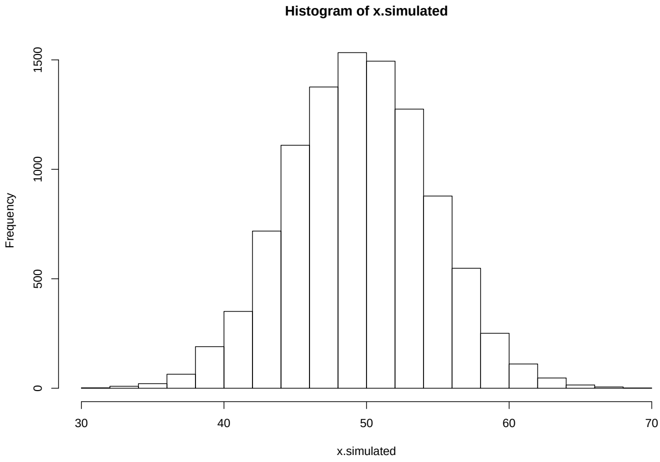
```
print(x.simulated[1:10])
```

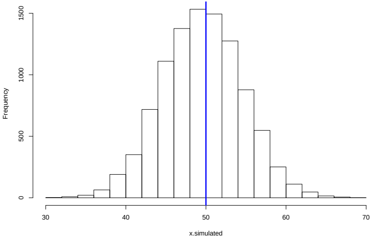
```
##  [1] 53 53 47 57 42 51 47 52 51 52
```

So, it is possible to obtain $x = 57$ heads even if the coin is fair!

Displays the empirical distribution of the number of heads, $x_{simulated}$, obtained from the simulation

```
hist(x.simulated)
```



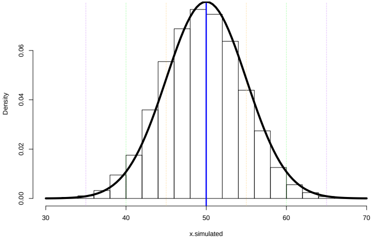



```
summary(x.simulated) # shows the range
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   47.00   50.00   50.01   53.00   69.00
```

```
summary(x.simulated)[4] # the mean
```

```
##    Mean
## 50.0107
```



The empirical sample mean is indeed very close to the theoretical expectation, $n \cdot \theta_0 = 100 \cdot 0.5 = 50$

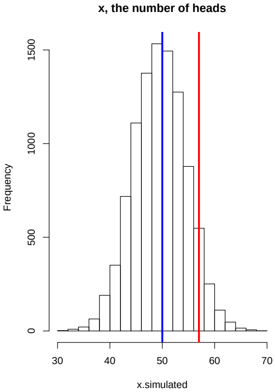
BUT, we can see there is some variation in the $x's$, in fact the 68%-95%-99.7% (area-under-curve) rule:

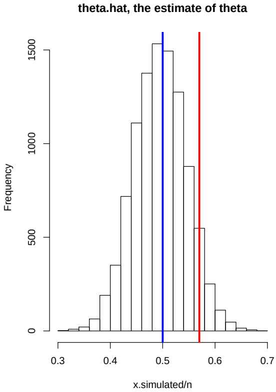
- 68% of the $x's$ fall between $\mu - \sigma = 50 - 5 = 45$ and $\mu + \sigma = 50 + 5 = 55$, $[45, 55]$
- 95% fall between $\mu - 2\sigma$ and $\mu + 2\sigma$, $[40, 60]$
- 99.7% fall between $\mu - 3\sigma$ and $\mu + 3\sigma$, $[35, 65]$

(Almost) Back to Square One

A coin is tossed $n = 100$ times and $x = 57$ heads are observed.

Is this a fair coin?





What is a p-value? (Wiki, February 8, 2021)

*In null hypothesis significance testing, the p-value is the probability of obtaining test results **at least as extreme** as the results actually observed, under the assumption that the null hypothesis is correct.*

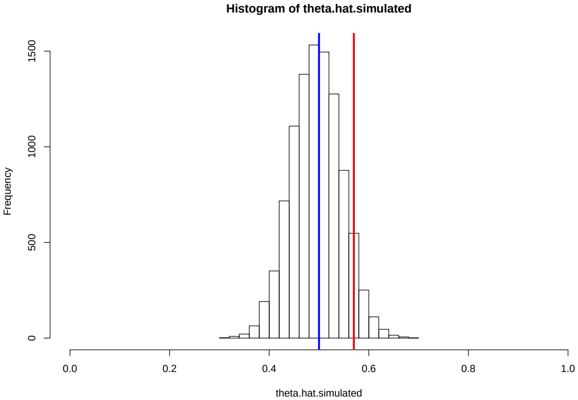
A very small p-value means that such an extreme observed outcome would be very unlikely under the null hypothesis.

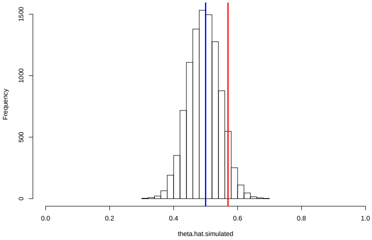
Reporting p-values of statistical tests is common practice in academic publications of many quantitative fields.

Since the precise meaning of p-value is hard to grasp, misuse is widespread and has been a major topic in metascience.

Simulation-based p-value Estimation, Emprical p-value

```
set.seed(1234)
n=100; x=57; theta.0=0.5 # the sample size, observed data, null hypothesis=fair coin
theta.hat=x/n # the point estimate of the parameter based on the observed data
n.rep=10000 # the number of experiments
x.simulated=rbinom(n.rep,n,theta.0) # Draws x's using a fair coin
theta.hat.simulated=x.simulated/n # Calcuates the corresponding theta estmiates
hist(theta.hat.simulated,xlim=c(0,1)) # Displays all the estimates
abline(v=theta.0, col="blue", lwd=3) # Marks theta.0
abline(v=theta.hat, col="red", lwd=3) # Marks theta.hat inferred from the actual observed data
```





- The above simulation will allow us to determine the p-value empirically (without knowing the formula).
- The p-value is the probability of obtaining test results (theta.hat.simulated), which were obtained under the null hypothesis (theta.0), as extreme as the observed result (theta.hat), which was obtained from the original experiment where we do not know the true value of theta.
- The rule of thumb is that if the p-value is less than 0.05 (5%), then it is unlikely the hypothesis that theta=theta.0 is true.
- We would declare that we reject the null hypothesis. In other words, the hypothesis test was statistically significant.

Calculating the Emprical p-value

We need to first determine how many theta.hat.simulated simulated from a fair coin (based on theta.0) are more extreme (bigger) than the observed value (theta.hat)

```
sum(theta.hat.simulated>=theta.hat)
```

```
## [1] 978
```

We then need to put this count into context with the total number of simulations This only takes into account one side of the histogram

```
sum(theta.hat.simulated>=theta.hat)/n.rep
```

```
## [1] 0.0978
```

Assuming that the distribution is symmetrical, we can double this value and that will be the "two-sided p-value"

```
2*sum(theta.hat.simulated>=theta.hat)/n.rep
```

```
## [1] 0.1956
```

What if x was 43?

The observed $\hat{\theta}$ would be smaller than θ_0, and we should count the replicates from the left tail.

Make the code more robust to this sort of changes

```
if (theta.hat>=theta.0){
    2*sum(theta.hat.simulated>=theta.hat)/n.rep
}else{
    2*sum(theta.hat.simulated<=theta.hat)/n.rep
}
```

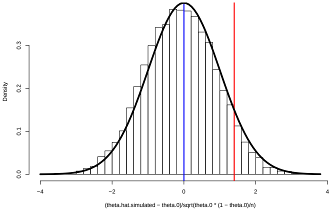
```
## [1] 0.1956
```

Looking ahead

Compare the empirial p-value to theoretical p-value calculated based on asymptotic Normal approximation

```
z.obs=(theta.hat-theta.0)/sqrt(theta.0*(1-theta.0)/n)
2*pnorm(abs(z.obs),lower.tail=F)
```

```
## [1] 0.1615133
```



The standardized test statistic

$$T = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{\theta_0 \cdot (1-\theta_0)}{n}}}$$

The distribution of T under the null hypothesis H_0

$$T \overset{H_0}{\sim} N(0,1) (= Z)$$

The asymptotic p-value calculation

$$\begin{aligned}
\text{p-value} &= 2 \cdot \text{Prob}(Z > z_{obs}) \\
&= 2 \cdot \text{Prob}(Z > \frac{0.57 - 0.5}{\sqrt{\frac{0.5 \cdot (1-0.5)}{100}}}) \\
&= 2 \cdot \text{Prob}(Z > 1.4) \\
&= 2 \cdot 0.08 = 0.16
\end{aligned}$$

The Pearson's χ^2 test

$$T = \sum_{k_{th} \text{ group}=1}^{K} = \frac{(O_k - E_k)^2}{E_k} \overset{H_0}{\sim} \chi^2_{K-1}.$$

The coin example

$$T = \frac{(57 - 100 * 0.5)^2}{100 * 0.5} + \frac{(43 - 100 * 0.5)^2}{43 * 0.5} = \frac{7^2}{50} + \frac{(-7)^2}{50} = 1.96$$

The asymptotic p-value calculation

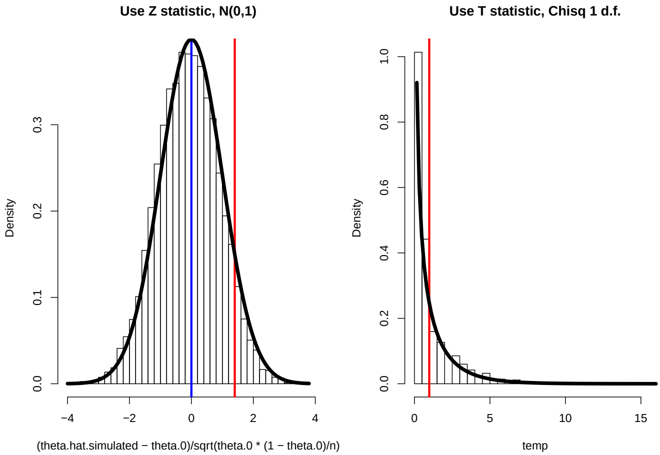
$$\text{p-value} = \text{Prob}(T > t_{obs}) = \text{Prob}(T > 1.96) = 0.16$$

```
1 - pchisq(1.96, df=1)
```

```
## [1] 0.1615133
```

Relationship between the two tests applied to the same data

Normal distribution vs. Chi-square distribution



Looking ahead, e.g. $Z^2 = \chi_1^2$. In the HWE testing problem, there were 3 groups, so why it was χ_1^2?

Answering the Question

A coin is tossed $n = 100$ times and $x = 57$ heads are observed.

Is this a fair coin?
$H_0 : \theta = \theta_0 = 0.5$

Because the p-value is > 0.05, we can "statistically" declare that, even though $\hat{\theta} = 0.57$ is mathematically different from $\theta_0 = 0.5$, we cannot confidently say that the original experiment did not use a fair coin, i.e.

The hypothesis test, testing H_0, was not statistically significant!

Food for Thoughts

- Instead of studying the number of heads, x, in n coin tosses which follows a discrete Binomial distribution, $Binom(n, \theta)$, can be study for example, height, which follows a continuous $N(\mu, \sigma^2)$ distribution?
- Understanding the effect of sample size on p-value

  (point estimation $\neq$ hypothesis testing)

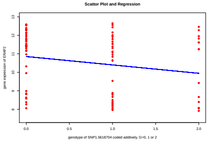
  A coin is tossed $n = 1000$ times and $x = 570$ heads are observed.

  $$\hat{\theta} = \frac{570}{1000} = 0.57$$

  Is this a fair coin?

- What is a false positive, and what is a false negative?
- What if we have a bag/family of 10^6 coins to evaluate?
- **How are these related with genetic association studies**?

Recall



The slope (the regression coefficient) is -0.4545. The slope is not statistically different from zero: the **p-value of testing the slope = 0 is 0.0594, not statistically significant.**

```
summary(lm(y~x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q  Median      3Q     Max
## -2.7969 -1.7987  0.5538  1.3051  2.5135
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.8544     0.2445  44.402   <2e-16 ***
## x            -0.4545     0.2380  -1.909   0.0594 .
## ---
```

What's next

How to use simulation to obtain the emprical p-value for

the association testing between the gene expression of $ERAP2$ (Y) and the genotypes of SNP1.5618704 coded additively.(X)

$$\text{Expected or average value of } Y = \beta_0 + \beta X.$$

That is, determine if the slope is zero, $H_0 : \beta = 0$.

What if we have a bag/family of 10^6 coins/SNPs to evaluate?