Introduction to Statistical Hypothesis Testing

Paul Cohen ISTA 370

January, 2012
Hypothesis Testing

Kinds of Hypothesis Testing

- Hypothesis: More women than men enroll at UA.
  - **Test 1**: Count all the women and men at UA and compare the numbers.
  - **Test 2**: Count all the women and men walking past the Canyon Cafe between noon and 1pm on five consecutive days and compare the numbers.

- How do these differ?
Kinds of Hypothesis Testing

- Hypothesis: More women than men enroll at UA.
  - **Test 1**: Count all the women and men at UA and compare the numbers.
  - **Test 2**: Count all the women and men walking past the Canyon Cafe between noon and 1pm on five consecutive days and compare the numbers.

- How do these differ?

- Hypothesis testing is for those cases in which an experiment result could have been different.

- Hypothesis testing is for inferring population parameters given sample statistics.
How to test the hypothesis that a coin is fair

Toss the coin 30 times. If it lands heads approximately 15 times, then it is probably fair. If does not land heads approximately 15 times, then it probably is not fair. **Why????**

N_h : number of heads in 30 tosses
How to test the hypothesis that a coin is fair

Toss the coin 30 times. If it lands heads approximately 15 times, then it is probably fair. If it does not land heads approximately 15 times, then it probably is not fair. **Why????**

Because if the coin were fair, then a very large or very small number of heads would be improbable.
The Logic of Hypothesis Testing

Form an hypothesis (the coin is fair). Do an experiment (toss the coin 30 times). If the hypothesis were true, then some experiment results would be improbable. The more improbable the result, the more confidence you can have if you reject the hypothesis.

N_h : number of heads in 30 tosses
Examples of Hypothesis Testing

What’s the hypothesis? What’s the experiment? If the hypothesis were true, what would be probable and improbable results?

- The efficacy of a cold-prevention drug
- The efficacy of a STEM program for middle-school girls
- The speed of a standard algorithm vs. a new one
Estimating the probability of experiment results

One experiment produces one result:

Results
Many *identical replications* of an experiment produce many results. Some are more frequent than others.
An infinite number of identical replications of an experiment produces a *probability distribution* of results (ask me how, later).
Start by defining some parameters of the coin-tossing experiment:

```r
> coin<-c('H','T')  # outcomes of tossing the coin
> probs<-c(0.5,0.5) # probabilities of H and T
> tosses<-30        # number of times we toss the coin
> sample(coin,tosses,replace=TRUE,probs)
```

```
[1] "H" "T" "T" "H" "H" "T" "H" "T" "H" "H" "T" "H" "T" "H" "H" "T" "H" "T" "H" "H"
[15] "H" "H" "H" "H" "T" "T" "H" "T" "H" "T" "H" "T" "H" "H" "H" "T" "H" "T" "H" "H"
[29] "H" "T"
```
Define a couple of functions to make the code clearer:

```r
> # Count the number heads in a sample s of tosses:
> CH<-function(s){length(which(s=="H"))}
> # Run a single coin-toss experiment:
> CoinTossExpt<-function(tosses,P){
>     CH(sample(coin,tosses,replace=TRUE, prob=P))
> }
```
Estimating the probability of number of heads, $N_h$

One experiment produces one result:

```r
> CoinTossExpt(tosses,probs)
[1] 17
```

Many experiments produce many results, some more likely than others:

```r
> sd<-replicate(50,CoinTossExpt(tosses,probs))
> table(sd) # shows values and their frequencies

sd
  7 11 12 13 14 15 16 17 18 21
 1 1  5  7 10  6  6  8  5  1
```
Estimating the probability of $N_h$

An infinite number of identical replications of an experiment produces a distribution of results.

```r
> k<-10000 # repeat the experiment k = `infinity' times
> samplingDist<-replicate(k,CoinTossExpt(tosses,probs))
> sd<-factor(samplingDist,levels=c(1:tosses)) # hack to fix R's
> sdRelativeFrequencies<-table(sd)
> plot(sdRelativeFrequencies)
```

These are frequencies of $N_h$ heads in 10000 replications of the experiment. We want probabilities, not frequencies.
R Tricks: Frequencies to Probabilities

Convert relative frequencies to probabilities by dividing each frequency by the sum of all frequencies

> # Sum the frequencies in a table of frequencies
> n<-sum(sdRelativeFrequencies)
> # Divide each frequency by n (n=k in this example)
> sdProbabilities<-(sdRelativeFrequencies / n)
> plot(sdProbabilities)
Review the logic...

Form an hypothesis (the coin is fair). Do an experiment (toss the coin 30 times). If the hypothesis were true, then some experiment results would be improbable. The more improbable the result, the less you believe the coin is fair.

\[ N_h : \text{number of heads in 30 tosses} \]

Now you know how to estimate the probability of experiment results!
If the hypothesis were true, some experiment results would be improbable.
If $N_h = 8$, would you say the coin is un-fair?

What if $N_h = 13$ or 19 or 27?
Suppose you toss a coin 30 times and it lands heads 20 times; is it a fair coin?

Hypothesis testing is like proof by contradiction. You always have two hypotheses: the one you want to prove and its converse.

- Assert the converse (called $H_0$, the *null hypothesis*)
- Use $H_0$ to find the *sampling distribution* of all possible experiment results.
- Perform an experiment, and if the result is very unlikely given $H_0$, then reject $H_0$ and “accept” the alternative $H_1$.

Once we know the sampling distribution of 30 tosses of a fair coin (why fair?) we can assess the probability that 20 of them are heads.
Hypothesis Testing Step by Step

1. State a *null hypothesis*, $H_0$: The coin is fair.
2. Perform an experiment: Toss the coin 30 times and get a *sample statistic*, $N_h$.
3. Find the *sampling distribution*, the probability distribution of the sample statistic, $N_h$, if $H_0$ were true.
4. Use the sampling distribution to calculate the probability, $p$, of your sample statistic, $N_h$, under the assumption that $H_0$ is true.
5. Based on $p$, decide whether to *reject* $H_0$. 
State a *null hypothesis*, $H_0$: The coin is fair.

More formally, $H_0 : P(Heads) = 0.5$.

```r
> coin<-c('H','T')  # outcomes of tossing the coin
> probs<-c(0.5,0.5)  # probabilities of H and T
```

The null hypothesis always states that a parameter has a *single value*, never a range, because this value is needed to find the sampling distribution.

Important geeky point: `probs<-c(.5,.5)` specifies just one value, because probabilities must sum to one.
Step 2: Run an Experiment

Perform an experiment: Toss the coin 30 times and get a sample statistic, \( N_h \).

Suppose you toss the coin 30 times and it comes up heads \( N_h = 20 \) times.

> # Number of heads = 20
> Nh<-20
Step 3: The Sampling Distribution
Define a Single Replication of the Experiment

Find the *sampling distribution*, the probability distribution of the sample statistic, $N_h$, if $H_0$ were true.

```r
> # Count the number heads in a sample s of tosses:
> CH<-function(s){length(which(s=="H"))}
> # Run a single coin-toss experiment:
> CoinTossExpt<-function(tosses,P){
>     Nh<-CH(sample(coin,tosses,replace=TRUE, prob=P))
>     return(Nh)
> }
```
Step 3: The Sampling Distribution
Replicate the Experiment, Calculate Probabilities

Find the sampling distribution, the probability distribution of the sample statistic, $N_h$, if $H_0$ were true.

```r
k<-10000  # close enough to infinite for our purposes
samplingDist<-replicate(k,CoinTossExpt(tosses,probs))
#convert frequency distribution to probability distribution
sd<-factor(samplingDist,levels=c(1:tosses))
sdRelativeFrequencies<-table(sd)
sdProbabilities<-(sdRelativeFrequencies / k)
```
Step 3: The Sampling Distribution
Step 4: Calculate Probability of Sample Statistic

Use the sampling distribution to calculate the probability, $p$, of your sample statistic, $N_h = 20$, under the assumption that $H_0$ is true.

```r
> # the probability of Nh=20 in the sampling distribution
> p<-sdProbabilities[20]
> p

20
0.0283
```

So $P(N_h = 20) = p = 0.0283$
Step 5: Decide

Based on $p$, decide whether to reject $H_0$.

$P(N_h = 20) = p = 0.0283$, so do you think the coin is fair?

If you decide the coin isn’t fair, what would be the probability that you are wrong?
Step 5: Decide

Based on $p$, decide whether to reject $H_0$.

$P(N_h = 20) = p = 0.0283$, so do you think the coin is fair?

If you decide the coin isn’t fair, what would be the probability that you are wrong?

$p = 0.0283$.

This is called the $p$ value, the probability of incorrectly rejecting the null hypothesis.
Review of Terms: Match up the items

- 20 of 30 tosses landed heads
- The coin is fair
- 30 tosses of the coin
- The distribution of all possible sample results
- Uncertainty remaining after rejecting $H_0$

- Sampling distribution
- A sample or experiment
- Null hypothesis
- $p$ value
- Sample result

How you’d write it up: “To test the null hypothesis that the coin is fair, we conducted an experiment in which the coin was tossed 30 times. The coin landed heads 20 times. Based on a Monte Carlo sampling distribution, we reject the null hypothesis ($p \leq 0.0283$).”
We rejected \( H_0 \) because \( N_h = 20 \) was improbable (\( p = 0.0283 \)) given \( H_0 \). But wouldn’t we have also rejected \( H_0 \) if \( N_h > 20 \)?

If so, then the \( p \) value would need to be adjusted.

If you’d reject \( H_0 \) given any of several results, then the \( p \) value is the sum of probabilities of all those results.

\[
P(N_h = 20 \text{ or } N_h = 21 \text{ or } \ldots) = P(N_h = 20) + P(N_h = 21) + \ldots
\]
\textbf{R Tricks: Plotting Probabilities}

\begin{verbatim}
> plot(sdProbabilities)
> points(c(1:10,20:30),
> as.vector(c(sdProbabilities[1:10],sdProbabilities[20:30]),
> type="h",col="blue",lwd=5)
\end{verbatim}
Probabilities of Regions

\[
\text{sum}(\text{sdProbabilities}[1:10])
\]
[1] 0.0472

\[
\text{sum}(\text{sdProbabilities}[20:30])
\]
[1] 0.0509

\[
\text{sum}(\text{sdProbabilities}[c(1:10,20:30)])
\]
[1] 0.0981
\[ p = \text{sum}(\text{sdProbabilities}[\text{c}(1:10,20:30)]) \]

So if we decided to reject \( H_0 \) when \( N_h \leq 10 \) or \( N_h \geq 20 \), then our \( p \) value – the probability that we are rejecting \( H_0 \) incorrectly – would be quite high: \( p = 0.0981 \).
Finding the “Right-sized” Rejection Region

> p <- sum(sdProbabilities[c(1:8, 22:30)])

However, if we reject $H_0$ only if the results are more extreme – $N_h \leq 8$ or $N_h \geq 22$, then $p = 0.0156$. 
Hypothesis Testing Step by Step

The old way of saying it:

4 Use the sampling distribution under $H_0$ to calculate the probability, $p$, of your sample statistic, $N_h$.

5 Based on $p$, decide whether to reject $H_0$.

The new way:

4a Decide on $\alpha$, a maximum acceptable probability of incorrectly rejecting $H_0$.

4b Use the sampling distribution under $H_0$ to find critical values $c^+$ and $c^-$ such that $P(N_h \geq c^+) + P(N_h \leq c^-) \leq \alpha$.

5 If $N_h \geq c^+$ or $N_h \leq c^-$, reject $H_0$. 
How to Decide

Critical Values

Decide on $\alpha$, a maximum acceptable probability of incorrectly rejecting $H_0$. Use the sampling distribution under $H_0$ to find critical values $c^+$ and $c^-$ such that $P(N_h \geq c^+) + P(N_h \leq c^-) \leq \alpha$.

If we decide $\alpha = 0.05$, then the critical values are $N_h = 9$ and $N_h = 21$ because $P(N_h \leq 9) + P(N_h \geq 21) = 0.043 \leq \alpha$. 