Simulation: Probability Made Concrete
Simulation-Based Critical Values

BTRY/STSCI 4

Simulating Marginal and Conditional Distributions

Used to notation $P(A|B)$ for the probability of A *given* B.

-

$$P(B|A) = \frac{P(}{}$$

We'll explore the computational equivalents of t

An Example

Data on time between eruptions at the 'old faithful' geyser in Yellowstone National Park:

We can represent this as being (approximately) t

$$f(x) = \frac{p}{\sqrt{2\pi}\sigma_1} e^{-(x-\mu_1)/2\sigma_1^2} + \frac{}{\sqrt{2\pi}\sigma_2} \qquad ^2$$

(note that because each normal distribution integrates to 1, their weighted sum does)

but it would be nice to simulate this data.

Simulating Mixture Models

We represented the two-peaked distribution above as two re-scaled normal distributions.

But we can construct it by posing a hypothetical binary random varia

from

1.
2. If $Z = 1$ simulate X from $N(\mu_1, \sigma$
   from $N(\mu_2, \sigma_2)$


```
Z = rbinom(1,1,p)
if(Z){ X = rnorm(1,mean=mu1,sd=sig1) }
else{ X = rnorm(1,mean=mu2,sd=sig2) }
```

See code for simulation (and vectorization).

Simulation and Probability

To translate simulation scheme into probability

$$Z \sim B(p), \; X|Z=1 \sim N(\mu_1, \sigma_1^2) \; X|Z=0 \sim N(\mu_2, \sigma_2^2)$$

so we have defined X *conditional* on Z.

But w

marg

$$P(X) = P(X|Z=1)P(Z=1) + P(X\ Z=0)P(Z=0)$$

$$= pN(\mu_1, \sigma_1^2) + (1-p)N($$

yielding the density above.

- Useful way of generating random variables (we'll see others later).
- Good way to think about probability: marginal distribution is what you get when you drop the information in Z.

Simulation and Bayes Theorem

We might also like to know which component to assign a given observation to.

ie, we're looking for

$P(Z$

Not c

`x[i]`


```
# How many Z=1 with X in range
Num = sum(component == 1 & mixdat > a & mixdat <= b)
# How many X in range
Den = sum(mixdat > a & mixdat <= b)
Pz = Num/Den
```

Statistical Tests Made Concrete

From Lecture 3:

- α-level: if the null hypothesis were true, we would (mistakenly)
-
  - Run hypothesis test.
  - Repeat many times; proportion rejecte

But we can also use this to define hypothesis te

- Most tests reject for (test statistic $>$ critic
- But we need to choose the critical value. Also by simulation!

Critical Values for Tests

(See R script for Lecture 7)

Suppose we want to test a hypothesis H_0 using data $X_1, \ldots, X_n$:

- Choose a statistic $t(X_1, \ldots, X_n)$ that should be small when H_0 is true and large when H_0 is false.
- But how do we actually find t if we don't trust current theory?
  - Simulate $X_1, \ldots, X_n$ under H_0.
    - Evaluate $T = t(X_1, \ldots, X_n)$.
    - Repeat to get $T_1, \ldots, T_N$.
    - t^α given by the quantile of $T_1, \ldots, T_N$.
- Note: problematic if H_0 does not *completely* specify distribution of $X_1, \ldots, X_n$.

A Negative Binomial Simulation

Back to testing the mean of a negative binomial (see Lecture 3)

```
nsim = 25000
n = 30
p = 0.07
mu = (1-
t.va

for(
  X = rnbinom(n,1,p)
  t.vals[i] = sqrt(n)*abs( mean(X) - mu )/s
}
```



```
> t.crit = quantile(t.vals,0.95) # Simulation critical
2.288837                         # value

> qt(0.975,29)    # t-distribution critical value
[1] 2.04523
```

Vectorizing

Let's see how to vectorize this (R script for timing):

```
# Generate nsim data sets over rows.
XX = matrix(rbinom(n*nsim,1,p),nsim,n)

# Take t
mean

# Subtract mean
center.X = XX - matrix(mean.X,nsim,n,by

# Average squared deviation then square-r
sd.X = sqrt(  (center.X^2)%*%rep(1/(n-1),n) )

# Caclulate Statistic
t.vals =  sqrt(n)*abs(mean.X - mu)/sd.X
```

Testing Two Populations

What if H_0 is pretty vague?

- Two samples $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ from distributions F_X and F_Y respectively.

Opti

- Two-sample t-test: $|X - Y| / \quad [n_1 s^2 + n_2 s^2]/(n_1 + n_2)$.
- Rank-sum test.

But:

- t-test critical value if you don't trust asymptotics?
- How do we think about other relationships (correlations, regression, ...)?

Constructing a Null Distribution

Idea (also behind rank sum):

- If the X's and Y's are from the same distribution, their *labels* shouldn't matter.
- So, if we randomly mix up their labels, things shouldn't change
- Now randomly permute the labels.

X_1	$\cdots$	X_{n_1}	Y_1		Y
1	$\cdots$	1	2		

- Treat permuting the labels like generating new X's and Y's.
- Evaluate t-statistic on the permuted labels; this is the *permutation distribution*.
- Rank-sum test is exactly a permutation test.

An Example Data Set

Example `chickwts` data in R gives weight of chickens fed different diets.

We will focus on differences between linseed and soybean.


```
data(chickwts)

X = chic

x = X[X$feed=='linseed',1]
y = X[X$feed=='soybean',1]

> t.test(x,y)
data:  x and y
t = -1.3246, df = 23.63, p-value = 0.198
```


But we'd like to verify that p=value.

A Test Statistic

We'll define a function to take `X` and give us the t statistic back.

```
chick.t.test = function(X){
  x = X[
  y = X[
  re
}
```


Defining this function is overkill (but saves space n

We could also use output of `lm(weigh`

First we'll record the observed statistic

```
t.obs = chick.t.test(X)
```

Constructing a Null Distribution

The `sample(N)` function will randomly re-arrange `1:N`.

```
> sample(5)
[1] 5 4 2 1 3
```

Now r eed.

```
nper                    # Number o
t.pe
temp.X = X              # Store a version of X that we can
                        # change arou
for(i in 1:nperm){
  I = sample(nrow(X))   # Generate a random perm
  temp.X[,2] = X[I,2]
  t.perm[i] = chick.t.test(temp.X)
}
```

Assessing Significance

Now we can ask *Is the observed statistic much larger than the permutation distribution?*


```
> mean(t.perm>t.obs)
[1] 0.194
```


We ca
value

```
> quantile(t.perm,0.95)
2.087385
```


Compare to t-value

```
> qt(0.975,23.63)
[1] 2.06561
```

Some Philosophical Distinctions

Permutation distribution has a different data-generating model.

- Null hypothesis: X's and Y's generated according to the same distribution.
-

So wh

- Under H_0 all permutations of X's a
- We can condition on the values in t
- and probability of rejecting is 0.05.
- But this is true for whatever the values of the data happen to be.

Formally

- We use the *order statistics*

$$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n_1+n_2)}$$

these are the values of the X's and Y's placed in order.

- https://eduassistpro.github.io/

- The α-level is the expectation over probability of rejecting given Z's.

$$P(t(X, Y) > t^{\alpha}(Z)) = E_{X,Y} P(t(\quad | \quad)_{X,Y}[\alpha] = \alpha$$

- Formally, permutation distribution results from uniform distribution on all $(n_1 + n_2)!$ permutations of labels (too large, so we work with random samples).

More Generally

Ideas extend to test associations between quantities:

- Correlations between two continuous random variables.
- Regression of a response onto multiple covariates.
- Sam

- If X and Y (possibly multivariate) are related, permuting one (either!) breaks the relationship.
- If they're independent, permuting one ma permutations are equally likely).

Choice of test statistic can be important (does it distinguish what you think is going on?)

Another Example

Look at all feeds in `chickwts`; do they affect outcome weight?

We'll use the F statistic for the regression.


```
mod = lm(weight~feed,data=chickwts)
fstat.obs = summary(mod)$fstatistic[1]

fsta
temp
for(i in 1:nperm){
  temp.data$feed = chickwts$feed[sa                ]
  fstat.perm[i] =
    summary(lm(weight~feed,data                    tic[1]
}

mean(fstat.perm>fstat.obs)
```

Limitations

- Restricted to breaking relationships.
- No option to partially break relationships.
-
  -
  - Could permute just the $_{i1}$; but this also changes relationship between x_{i1} and $x_{i2} \Rightarrow$ changes v coefficients
  - Some variations to let you do this later.
- Not always the most powerful test available.
- **But**: pretty generic when applicable.

More General Statistics

Standard test statistics are not the only measures that can be permuted.

- Kolmogorov-Smirnoff test for two samples: maximal distance

- t also

- Could compare variances, if you think that this is the most obvious difference in distributions.

- Relationships between collections of cor
  10 ecological covariates and 4 human land-
  correlation, major canonical covariate.
- Little theory to guide best statistic; choice is based on what will pick up the signal you expect to find.

Summary

- Probability: often very helpful to think about theory via what simulation looks like.
-
-
- Permutation tests: randomly re-order so break up relationships in the data.
-
- Ie, make H_0 true; then use observed dat
- Next: multiple testing and false discovery rates.