Simulation: Probability Made Concrete

Simulation-Based Critical Values

Permutation Tests

Rizzo Chapter 8

BTRY/STSCI 4520

Simulating Marginal and Conditional Distributions

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

- Used for marginal distributions
  $P(A) = \sum_B P(A|B)P(B) = E_B P(A|B)$

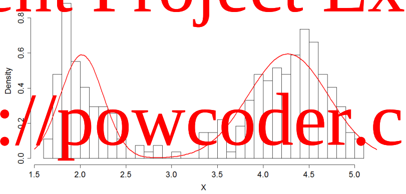
- In Bayes theorem:

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

We'll explore the computational equivalents of these here.

An Example

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



We can represent this as being (approximately) two normal peaks

$$f(x) = \frac{p}{\sqrt{2\pi}\sigma_1} e^{-(x-\mu_1)/2\sigma_1^2} + \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{-(x-\mu_2)/2\sigma_2^2}$$

(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 variable Z to tell us which normal component an observation comes from

1. Simulate Z as Bernoulli with probability p
2. If $Z = 1$ simulate X from $N(\mu_1, \sigma_1)$, otherwise simulate X 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

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

so we have defined X *conditional* on Z.

But when we look at X by itself (ie, throw away Z) we get the *marginal* distribution

$$\begin{aligned} P(X) &= P(X|Z = 1)P(Z = 1) + P(X|Z = 0)P(Z = 0) \\ &= pN(\mu_1, \sigma_1^2) + (1 - p)N(\mu_2, \sigma_2^2) \end{aligned}$$

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
$P(Z = 1|X = x) = P(X = x|Z = 1)P(Z = 1)/P(X = x)$

Not quite possible to do in data (usually only one example of `x[i]`), but we can look at $P(Z = 1|X \in [a,\ b])$.

```
# 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) reject α-proportion of the time.
- To check this
  - Simulate data generated from the null distribution.
  - Run hypothesis test.
  - Repeat many times; proportion rejected should be α.

But we can also use this to define hypothesis tests:

- Most tests reject for (test statistic $>$ critical value)
- 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.
- Reject H_0 if $t(X_1, \ldots, X_n) > t^\alpha$.
- Define t^α so that $P(t(X_1, \ldots, X_n) > t^\alpha | H_0) = \alpha$.
- But how do we actually find t^α 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-p)/p
t.vals = rep(0,nsim)

for(i in 1:nsim){         # Data and t-statistic
  X = rnbinom(n,1,p)
  t.vals[i] = sqrt(n)*abs( mean(X) - mu )/sd(X)
}

> 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 the mean
mean.X = XX%*%rep(1/n,n)

# Subtract mean
center.X = XX - matrix(mean.X,nsim,n,byrow=FALSE)

# Average squared deviation then square-root
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.
- $H_0 : F_X = F_Y$, but F_X not specified.

Options:

- Two-sample t-test: $|\bar{X} - \bar{Y}| / \sqrt{[n_1 s_X^2 + n_2 s_Y^2]/(n_1 + n_2)}$.
- 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 very much.
- Implementation: add a column (row below to save space) of labels

$$\begin{bmatrix} X_1 & \cdots & X_{n_1} & Y_1 & \cdots & Y_{n_2} \\ 1 & \cdots & 1 & 2 & \cdots & 2 \end{bmatrix}$$

  Now randomly permute the labels.
- 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 = chickwts[chickwts$feed=='linseed'|
                chickwts$feed=='soybean',]
```

```
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[X$feed=='linseed',1]  # Split into linseed
  y = X[X$feed=='soybean',1]  # and soybean
  return( abs(t.test(x,y)$statistic) )
}
```

Defining this function is overkill (but saves space next slide).

We could also use output of `lm(weights~feed,data=X)`

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 record the *t* statistic under random permutations of feed.

```
nperm = 1000               # Number of permutations
t.perm = rep(0,1000)
temp.X = X                 # Store a version of X that we can
                           # change around.

for(i in 1:nperm){
  I = sample(nrow(X))  # Generate a random permutation.
  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?*

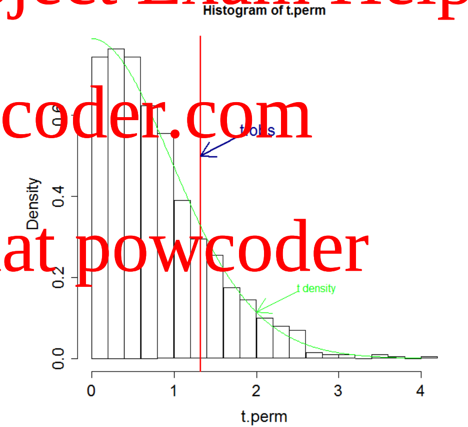
```
> mean(t.perm>t.obs)
[1] 0.194
```

We can also look at the critical 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.
- Permutation test: X's and Y's fixed, but labels are assigned at random.

So why is the permutation test OK?

- Under H_0 all permutations of X's and Y's are equally likely.
- We can condition on the *values* in the data set (not labels) 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.

- We can *condition* on these (ie, we got these values, whatever the labels). Critical value is a function of the Z's, probability of rejecting *given* Z's is α.

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

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

- 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.
- Associations between *groups* of covariates.

Same reasoning as before:

- If X and Y (possibly multivariate) are related, permuting one (either!) breaks the relationship.
- If they're independent, permuting one makes no difference (all 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]

fstat.perm = rep(0,nperm)
temp.data = chickwts
for(i in 1:nperm){
  temp.data$feed = chickwts$feed[sample(nrow(chickwts))]
  fstat.perm[i] =
    summary(lm(weight~feed,data=temp.data))$fstatistic[1]
}

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

Limitations

- Restricted to breaking relationships.
- No option to partially break relationships.
- Eg: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
  - Can test $\beta_1 = \beta_2 = 0$ by permuting the y_i.
  - Suppose we wanted to test just $\beta_1 = 0$? Could permute just the x_{i1}; but this also changes relationship between x_{i1} and x_{i2} ⇒ changes variance of your estimated 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 between empirical cdfs.
- Comparing two multivariate samples: Hotelling's T^2, but also other measures (Rizzo's own distance covariance).

- Could compare variances, if you think that this is the most obvious difference in distributions.
- Relationships between collections of continuous covariates (eg 10 ecological covariates and 4 human land-use): largest 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.
- For conducting tests; when in doubt, simulate!

- In R, clever vectorization can buy you a lot of speed (when you need it).
- Permutation tests: randomly re-order some columns to break-up relationships in the data.

- Ie, make H_0 true; then use observed data to conduct your test.
- Next: multiple testing and false discovery rates.