PQHS 471
Lecture 12: Housekeeping Utilities in Statistical Learning

Simulation: Random Number Generation, Permutation

Statistical Simulation

- Statistical simulation (Monte Carlo) is an important part of statistical method research.
- The statistical theories/methods are all based on assumptions. So most theorems state something like "if the data follow these models/assumptions, then...".
- The theories can hardly be verified in real world data because (1) the real data never satisfy the assumption; and (2) the underlying truth is unknown (no "gold standard").
- In simulation, data are "created" in a well controlled environment (model assumptions) and all truth are known. So the claim in the theorem can be verified.

Random Number Generator (RNG)

- Random number generator is the basis of statistical simulation. It serves to generate random numbers from predefined statistical distributions.
- Traditional methods (flip a coin or dice) work, but can't scale up.
- Computational methods are available to generate "pseudorandom" numbers.

Random Number Generator (RNG)

The random number generation often starts from generating uniform(0,1).
The most common method: "**Linear congruential generator**":

$$X_{n+1} = (aX_n + c) \text{ mod } m$$

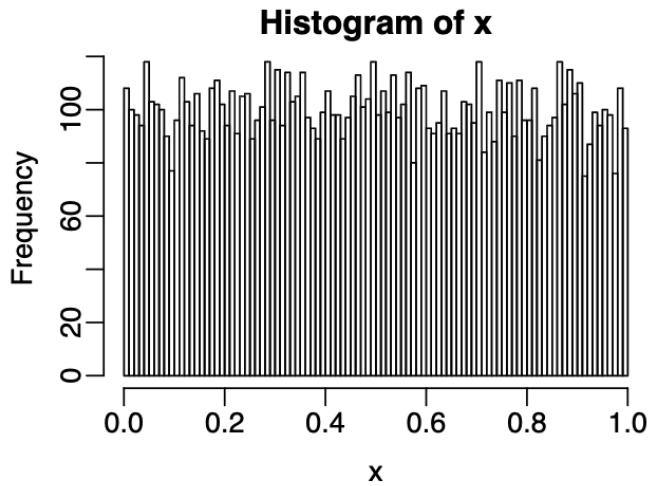
Here, a, c, and m are predefined numbers:

- X_0: random number "seed".
- a: multiplier, 1103515245 in *glibc*.
- c: increment, 12345 in *glibc*.
- m: modulus, 2^{32} or 2^{64}.

$U_n = X_n/m$ is distributed as Uniform(0,1).

Linear congruential generator

```
a = 1103515245; c = 12345; m = 2^32
n = 10000
x = numeric(n)
x[1] = 1
for( i in 2:n) {
x[i] = (a*x[i-1] + c) %% m
}
x = x/m
hist(x, 100)
```



Random Number Generator (RNG)

A few remarks about Linear congruential generator:

- The numbers generated will be exactly the same using the same seed.
- Want cycle of generator (number of steps before it begins repeating) to be large.
- Don't generate more than m/1000 numbers.

RNG in *R*:

- *set.seed* is the function to specify random seed.
- Read the help for *.Random.seed* for more description about random number generation in *R*.
- *runif* is used to generate uniform(0,1) r.v.

My recommendation: always set and save random number seed during simulation, so that the simulation results can be reproduced.

Simulate r.v. from other distributions

When the distribution has a **cumulative distribution function (cdf)** $\mathbf{F}$, the r.v. can be obtained by inverting the cdf ("inversion sampling"). This is based on the theory that the cdf is distributed as Uniform (0,1):
Algorithm: Assume $\mathbf{F}$ is the cdf of distribution $\mathcal{D}$. Given $u \sim unif(0,1)$, find a unique real number x such that $\mathbf{F}(x) = u$. Then $x \sim \mathcal{D}$.

Simulate r.v. from other distributions

When the distribution has a **cumulative distribution function (cdf)** $\mathbf{F}$, the r.v. can be obtained by inverting the cdf ("inversion sampling"). This is based on the theory that the cdf is distributed as Uniform (0,1):
Algorithm: Assume $\mathbf{F}$ is the cdf of distribution $\mathcal{D}$. Given $u \sim unif(0,1)$, find a unique real number x such that $\mathbf{F}(x) = u$. Then $x \sim \mathcal{D}$.
Example: exponential distribution. When $x \sim exp(\lambda)$, the cdf is $\mathbf{F}(x) = 1 - exp(-\lambda x)$. The inversion of cdf is:

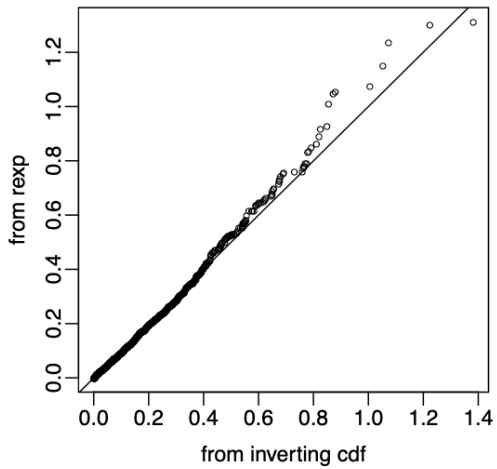
$$\mathbf{F}^{-1}(\mu) = -log(1-\mu)/\lambda$$

Then to generate exponential r.v., do:

- Generate uniform(0,1) r.v. denoted by μ.
- Calculate $x = -log(1-\mu)/\lambda$

Example: simulate exponential r.v.

```
lambda=5
u = runif(1000)
x = -log(1-u) / lambda
## generate from R's function
x2 = rexp(1000, lambda)
## compare
qqplot(x, x2, xlab="from inverting cdf", ylab="from rexp")
abline(0,1)
```



Simulate random vectors

Difficulty: Generating random vectors is more difficult, because we need to consider the correlation structure.
Solution: Generate **independent** r.v.'s, then apply some kind of transformation.
Example: simulate from multivariate normal distribution $MVN(\mu, \Sigma)$
Let Z be a p-vector of independent $N(0,1)$ r.v.'s, given $p \times p$ matrix D:

$$var(D^T Z) = D^T var(Z) D = D^T D$$

Therefore, the simulation steps are:

1. Perform **Cholesky decomposition** on Σ to find D s.t. $\Sigma = D^T D$.
2. Simulate $Z = (z_1, z_2, ..., z_p)' \sim N(0,1)$
3. Apply transformation $X = D^T Z + \mu$

R function mvrnorm available in *MASS* package.

Example: generate multivariate normal

```
## specify mean and variance/covariance matrix
mu = c(0,1)
Sigma = matrix(c(1.7, 0.5, 0.5, 0.8), nrow=2)
## Cholesky decomposition
D = chol(Sigma)
## generate 500 Z's.
Z = matrix(rnorm(1000), nrow=2)
## transform
X = t(D) %*% Z + mu
## check the means X
> rowMeans(X)
[1] -0.08976896  0.95802769
## check the variance/covariance matrix of X
> cov(t(X))
[,1]      [,2]
[1,] 1.7392114 0.5609027
[2,] 0.5609027 0.7380548
```

Permutation

- In statistical inference, it is important to know the distribution of some statistics under null hypothesis (H_0), so that quantities like p-values can be derived.
- The null distribution is available theoretically in some cases. For example,assume $X_i \sim N(\mu, \sigma^2), i = 1, ..., n$. Under $H_0 : \mu = 0$, we have $\bar{X} \sim N(0, \sigma^2/n)$. Then H_0 can be tested by comparing $\bar{X}$ with $N(0, \sigma^2/n)$.
- When null distribution cannot be obtained, it is useful to use **permutation test** to "create" a null distribution from data.

Permutation

The basic procedure of permutation test for H_0:

- Permute data under H_0 for a number of times. Each time recompute the test statistics. The test statistics obtained from the permuted data form the null distribution.
- Compare the observed test statistics with the null distribution to obtain statistical significance.

Permutation test example

Assume there are two sets of independent normal r.v.'s with the same known variance and unknown means: $X_i \sim N(\mu_1, \sigma^2), Y_i \sim N(\mu_2, \sigma^2)$.
We wish to test $H_0 : \mu_1 = \mu_2$.
Define the test statistics: $t = \bar{X} - \bar{Y}$. We know under the null, we have $t \sim N(0, 2\sigma^2/n)$ (assuming same sample size n in both groups). Using the permutation test, we do:

1. Pool X and Y together, denote the pooled vector by Z.
2. Randomly shuffle Z. For each shuffling, take the first n items as the new X (denote as X^*) and the next n items as the new Y (denoted as Y^*).
3. Compute $t^* = \bar{X}^* - \bar{Y}^*$.
4. Repeat steps 2 and 3 for a number of times. The result t^*'s form the null distribution of t.
5. To compute p-values, calculate $Pr(|t^*| > |t|)$.

Note: the random shuffling is based on H_0, that X and Y are i.i.d.

Example: permutation test

```
> x=rnorm(100, 0, 1)
> y=rnorm(100, 0.5, 1)
> t.test(x,y)
Welch Two Sample t-test
data:  x and y
t = -1.9751, df = 197.962, p-value = 0.04965

> nsims=50000
> t.obs = mean(x) - mean(y)
> t.perm = rep(0, nsims)
> for(i in 1:nsims) {
+     tmp = sample(c(x,y))
+     t.perm[i] = mean(tmp[1:100]) - mean(tmp[101:200])
+ }
> mean(abs(t.obs) < abs(t.perm))
[1] 0.04814
```

Permutation test: regression example

- Under linear regression setting (without intercept) $y_i = \beta x_i + \epsilon_i$. We want to test the coefficient: $H_0 : \beta = 0$.
- Observed data are (x_i, y_i) pairs.
- Use ordinary least square estimator for β, denote as $\hat{\beta}(\mathbf{x}, \mathbf{y})$.

The permutation test steps are:

1. Keep y_i unchanged, permute (change the order of) x_i to obtain a vector, denoted as x_i^*.
2. Obtain estimate under the permuted data: $\hat{\beta}^*(\mathbf{x}^*, \mathbf{y})$.
3. Repeat step 1 and 2. $\hat{\beta}^*$'s form the null distribution for $\hat{\beta}$.
4. P-value $= Pr(|\hat{\beta}^*| > \hat{\beta})$.

Note: the random shuffling of x_i is based on the H_0, that is there is no association between $\mathbf{x}$ and $\mathbf{y}$.

Example: regression permutation test

```
> x = rnorm(100); y = 0.2 * x + rnorm(100)
> summary(lm(y~x-1))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
x   0.1502     0.1050   1.431    0.156
> nsims=5000
> beta.obs = coef(lm(y~x-1))
> beta.perm = rep(0, nsims)
> for(i in 1:nsims) {
+     xstar = sample(x)
+     beta.perm[i] = coef(lm(y~xstar-1))
+ }
> mean(abs(beta.obs) < abs(beta.perm))
[1] 0.157
```

Good Statistical Practice

The Lady Tasting Tea

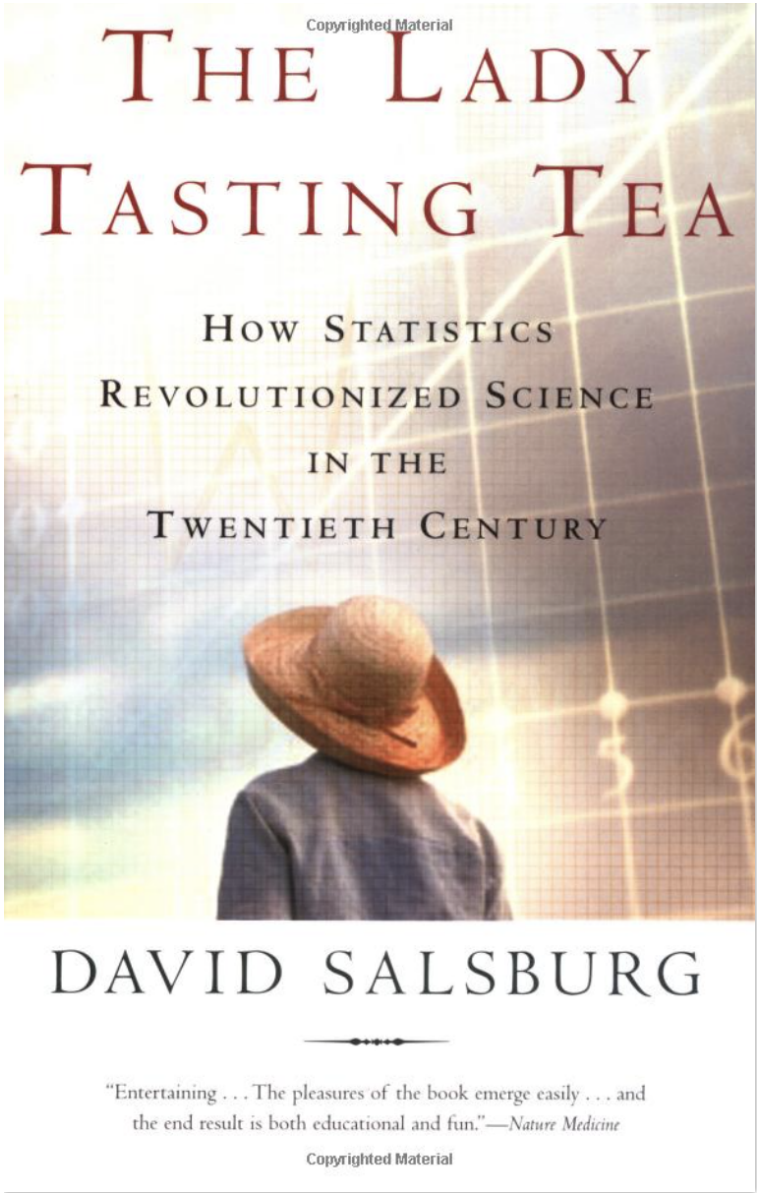
- It was a summer afternoon in Cambridge, England, in the 1920s.
- A groups of university dons, their wifes, and some guests were having afternoon tea.
- A lady was insisting that tea tasted different depending upon whether *the tea was poured into the milk* OR *the milk was poured into the tea*.

R.A. Fisher

Fisher in 1913

- "Sheer nonsense", the scientific minds among the men scoffed at this.
- A thin, short man, with thick glasses, Ronald Fisher, pounced on the problem: "Let us test the proposition!"

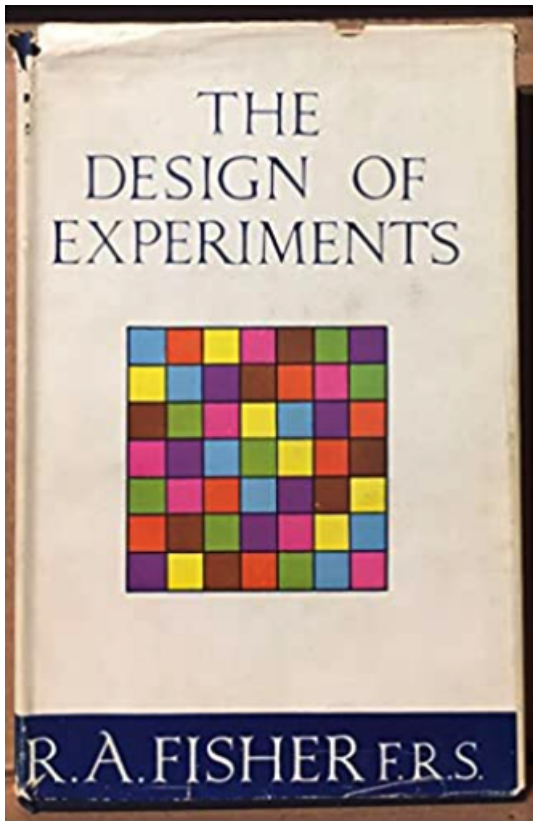
The Lady Tasting Tea



Hypothesis Testing

- Fisher's notion of a *null hypothesis*
  — Null hypothesis
  — Popularize p-value
- Neyman-Pearson Lemma
  — Error of the 2nd kind
  — Alternative/competing hypothesis
  — Power function

Most influential books on statistical methods



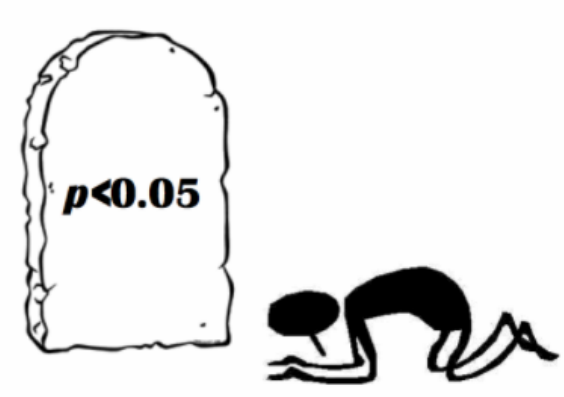
- **Statistical Methods for Research Workers**
- **The Design of Experiments**

“...the best thing about being a statistician...”

John Wilder Tukey

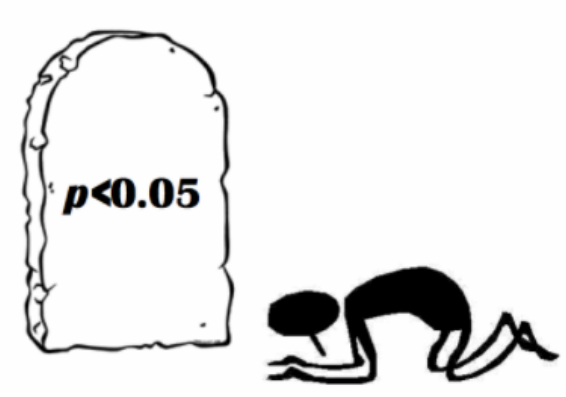
“... is that you get to play in everyone’s backyard.”

Misuse of p-value



- Q: Why do so many colleges and grad schools teach p = 0.05?
- A: Because that's still what the scientific community and journal editors use.
- Q: Why do so many people still use p = 0.05?
- A: Because that's what they were taught in college or grad school.

Misuse of p-value



- Q: Why do so many colleges and grad schools teach p = 0.05?
- A: Because that's still what the scientific community and journal editors use.
- Q: Why do so many people still use p = 0.05?
- A: Because that's what they were taught in college or grad school.

"We teach it because it's what we do; we do it because it's what we teach."

Fisher's words in SMRW

"Personally, the writer prefers to set a low standard of significance at 5 percentage point... A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance."

ASA Statement on p-values

Taylor & Francis

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

The ASA Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA Statement on *p*-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

To link to this article: https://doi.org/10.1080/00031305.2016.1154108

pop quiz

Which(s) of the following statements is/are reasonable?

- p-value is a probability.
- $p > 0.05$ is the probability that the null hypothesis is true.
- 1 minus the p-value is the probability that the alternative hypothesis is true.
- A statistically significant test result ($p \leq 0.05$) means that the test hypothesis is false or should be rejected.
- A p-value greater than 0.05 means that no effect was observed.

The status quo

Informally, a p-value is the probability **under a specified statistical model** that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be *equal to or more extreme* than its observed value.

Six principles of p-value

- 1. P-values can indicate how incompatible the data are with a specified statistical model.
  — The most common context is a model (under a set of assumptions): H_0
  — Often H_0 postulates the absence of an effect (e.g. no difference between two groups)
  — The smaller the p-value, the greater the incompatibility of the data with H_0
  — Incompatibility casting doubt on H_0

Six principles of p-value

- 1. P-values can indicate how incompatible the data are with a specified statistical model.
  — The most common context is a model (under a set of assumptions): H_0
  — Often H_0 postulates the absence of an effect (e.g. no difference between two groups)
  — The smaller the p-value, the greater the incompatibility of the data with H_0
  — Incompatibility casting doubt on H_0
- 2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
  — Never turn a p-value into a statement about the truth of H_0
  — p-value is a statement about the **relationship** between the data and H_0, NOT about the **explanation** (H_0) itself.

Six principles of p-value (cont'd)

- 3. Scientific conclusions and business or policy decisions should NOT be based only on whether a p-value passes a specific threshold.
  — "bright-line" rule (e.g. $p < 0.05$ alone) can lead to erroneous beliefs and poor decision making.
  — A conclusion does not immediately become "true" on one side of the divide and "false" on the other.
  — Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis.
  — Using $p < 0.05$ alone as a license for making a claim of a scientific finding leads to considerable distortion of the scientific process.

Six principles of p-value (cont'd)

- 3. Scientific conclusions and business or policy decisions should NOT be based only on whether a p-value passes a specific threshold.
  — "bright-line" rule (e.g. $p < 0.05$ alone) can lead to erroneous beliefs and poor decision making.
  — A conclusion does not immediately become "true" on one side of the divide and "false" on the other.
  — Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis.
  — Using $p < 0.05$ alone as a license for making a claim of a scientific finding leads to considerable distortion of the scientific process.
- 4. Proper inference requires full reporting and transparency
  — number of hypotheses explored, all data collection decisions, all statistical analyses conducted
  — No "cherry-picking"

Six principles of p-value (cont'd)

- 5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
  — $pval \neq$effect size
  — Statistical sig. vs. biological sig.

Six principles of p-value (cont'd)

- 5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
  — $pval \neq$effect size
  — Statistical sig. vs. biological sig.
- 6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Usage of p-value

- **Good statistical practice** is an integral part of **good scientific practice**.
  — study design and conduct, summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting, proper logical understanding of results.

Usage of p-value

- **Good statistical practice** is an integral part of **good scientific practice**.
  — study design and conduct, summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting, proper logical understanding of results.
- **No single index should substitute for scientific reasoning.**