Permutation and Randomization Tests

Mark M. Fredrickson (`mfredric@umich.edu`)

Computational Methods in Statistics and Data Science (Stats 406)

Permutation Tests

Permutation invariant test statistics

When discussing **the bootstrap**, we wanted estimate a parameter θ using,

$$\hat{\theta} = T(X_1, \ldots, X_n)$$

Permutation invariant test statistics

When discussing **the bootstrap**, we wanted estimate a parameter θ using,

$$\hat{\theta} = T(X_1, \ldots, X_n)$$

To get the distribution of T, we **conditioned** on the sample $X_1, \ldots, X_n$ and generated samples (with replacement) from the observed values.

Permutation invariant test statistics

When discussing **the bootstrap**, we wanted estimate a parameter θ using,

$$\hat{\theta} = T(X_1, \ldots, X_n)$$

To get the distribution of T, we **conditioned** on the sample $X_1, \ldots, X_n$ and generated samples (with replacement) from the observed values.

Permutation tests also **condition on aspects of the sample** such that the remaining data is **invariant to permutation**.

Permutation invariant test statistics

When discussing **the bootstrap**, we wanted estimate a parameter θ using,

$$\hat{\theta} = T(X_1, \ldots, X_n)$$

To get the distribution of T, we **conditioned** on the sample $X_1, \ldots, X_n$ and generated samples (with replacement) from the observed values.

Permutation tests also **condition on aspects of the sample** such that the remaining data is **invariant to permutation**.

A close analog: assume $Y = \beta_0 + \beta_1 X + \epsilon$ and condition on $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The $\hat{\epsilon} = y - \hat{y}$ is **permutation invariant** (we can shuffle them around freely).

Permutation tests

We will consider a variety of **hypothesis tests** that exploit invariance. Suppose that Z is invariant conditional on the sample, then we need:

$$T(Z_1, \ldots, Z_n) \sim T(\pi(Z_1, \ldots, Z_n))$$

where π is a permutation of the Z_i values (shuffle).

Permutation tests

We will consider a variety of **hypothesis tests** that exploit invariance. Suppose that Z is invariant conditional on the sample, then we need:

$$T(Z_1, \ldots, Z_n) \sim T(\pi(Z_1, \ldots, Z_n))$$

where π is a permutation of the Z_i values (shuffle).

Then to get the null distribution, we **permute the sample** and compute T.

Permutation tests

We will consider a variety of **hypothesis tests** that exploit invariance. Suppose that Z is invariant conditional on the sample, then we need:

$$T(Z_1, \ldots, Z_n) \sim T(\pi(Z_1, \ldots, Z_n))$$

where π is a permutation of the Z_i values (shuffle).

Then to get the null distribution, we **permute the sample** and compute T.

Tests of this form are known as **permutation tests** when **sampling** and **randomization tests** when analyzing **randomized controlled trials**.

Example: Developing a test for symmetry

Suppose we have some data from

$$X_i \overset{\text{iid}}{\sim} F, -\infty < x < \infty, \text{continuous}$$

and we want to test the hypothesis:

$$H_0 : P(X \leq \theta_0) = 0.5, P(X \leq \theta_0 - t) = P(X \geq \theta + t)$$

Example: Developing a test for symmetry

Suppose we have some data from

$$X_i \overset{\text{iid}}{\sim} F, -\infty < x < \infty, \text{continuous}$$

and we want to test the hypothesis:

$$H_0 : P(X \leq \theta_0) = 0.5, P(X \leq \theta_0 - t) = P(X \geq \theta + t)$$

In other words, X is **symmetric about θ_0**.

Example: Developing a test for symmetry

Suppose we have some data from

$$X_i \overset{\text{iid}}{\sim} F, -\infty < x < \infty, \text{continuous}$$

and we want to test the hypothesis:

$$H_0 : P(X \leq \theta_0) = 0.5, P(X \leq \theta_0 - t) = P(X \geq \theta + t)$$

In other words, X is **symmetric about θ_0**.

We'll develop a test in two parts: (a) a test for medians, and (b) a permutation test for symmetry

Sign test

Let θ be the **median of** F: $P(X \leq \theta) = 0.5$.

Sign test

Let θ be the **median of F**: $P(X \leq \theta) = 0.5$.

As we've seen, the random variable $Y = I(X \leq \theta)$ is a **Bernoulli**(0.5) and

$$S = \sum_{i=1}^{n} Y_i \sim \text{Binomial}(n, 0.5)$$

Sign test

Let θ be the **median of F**: $P(X \leq \theta) = 0.5$.

As we've seen, the random variable $Y = I(X \leq \theta)$ is a **Bernoulli**(0.5) and

$$S = \sum_{i=1}^{n} Y_i \sim \text{Binomial}(n, 0.5)$$

Under the null hypothesis $H_0 : \theta = \theta_0$, we have a **null distribution** $S \sim Binomial(n, 0.5)$.

Let θ be the **median of F**: $P(X \leq \theta) = 0.5$.

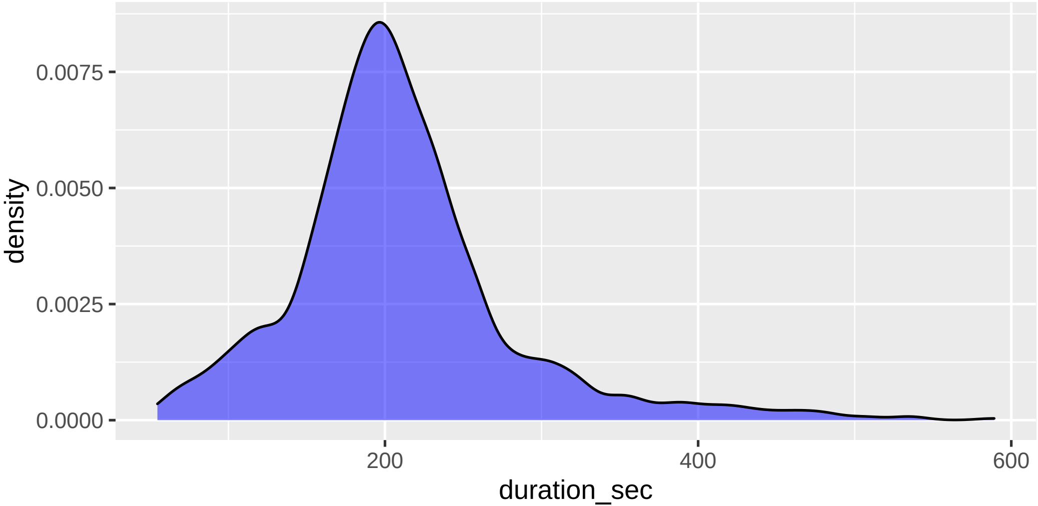
As we've seen, the random variable $Y = I(X \leq \theta)$ is a **Bernoulli**(0.5) and

$$S = \sum_{i=1}^{n} Y_i \sim \text{Binomial}(n, 0.5)$$

Under the null hypothesis $H_0 : \theta = \theta_0$, we have a **null distribution** $S \sim Binomial(n, 0.5)$.

We can use `binom.test` to test this hypothesis.

Application: Length of songs on Spotify less than 10 minutes



Testing $\theta = 200$

```
> testMedian <- function(median0) {
+     y <- tracks10$duration_sec - median0 > 0
+     binom.test(sum(y), length(y), p = 0.5)$p.value
+ }
> testMedian(200)

[1] 0.385
```

Testing $\theta = 200$

```
> testMedian <- function(median0) {
+     y <- tracks10$duration_sec - median0 > 0
+     binom.test(sum(y), length(y), p = 0.5)$p.value
+ }
> testMedian(200)
```

```
[1] 0.385
```

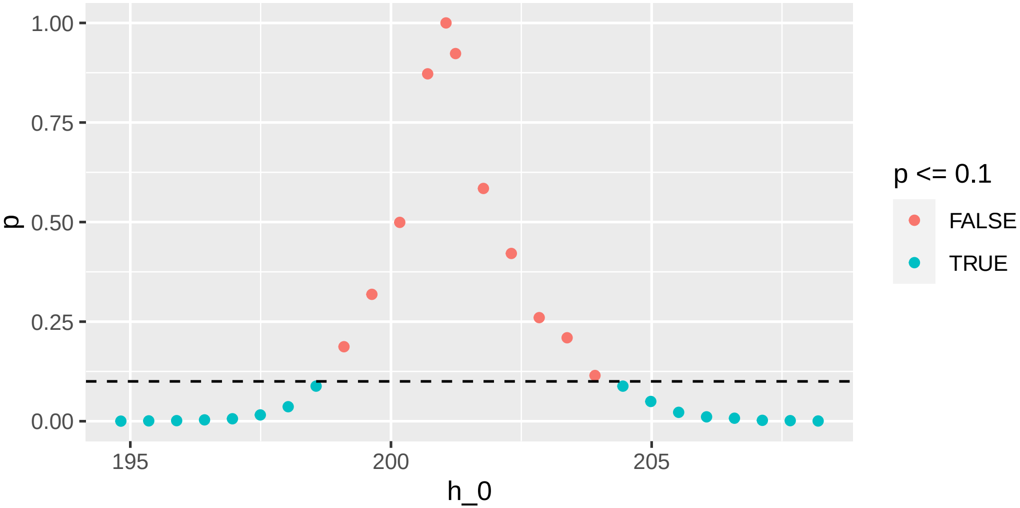
Confidence intervals

```
> medians <- unique(sort(c(
+               median(tracks10$duration_sec),
+               seq(min(tracks10$duration_sec),
+                 max(tracks10$duration_sec),
+                 length.out = 1000))))
> pvalues <- map_dbl(medians, testMedian)
```

Confidence intervals

```
> medians <- unique(sort(c(
+             median(tracks10$duration_sec),
+             seq(min(tracks10$duration_sec),
+               max(tracks10$duration_sec),
+               length.out = 1000))))
> pvalues <- map_dbl(medians, testMedian)

> medians[which.max(pvalues)] # point estimate

[1] 201.1

> range(medians[0.001 <= pvalues]) # 99.9% CI

[1] 195.9 207.7
```

Graphing p-values (90% CI)



Symmetry

Recall the **decomposition of Laplace/Double Exponential** random variables that were **symmetric about θ**:

$$X = SY + \theta$$

where S was the **sign of X**, Y was the **magnitude**.

Symmetry

Recall the **decomposition of Laplace/Double Exponential** random variables that were **symmetric about θ**:

$$X = SY + \theta$$

where S was the **sign of X**, Y was the **magnitude**.

Notice $Y = |X - \theta|$. If we **condition on Y**, then

$$P(X - \theta = y \mid Y = y) = P(X - \theta = -y \mid Y = y) = 0.5$$

Symmetry

Recall the **decomposition of Laplace/Double Exponential** random variables that were **symmetric about θ**:

$$X = SY + \theta$$

where S was the **sign of X**, Y was the **magnitude**.

Notice $Y = |X - \theta|$. If we **condition on Y**, then

$$P(X - \theta = y \mid Y = y) = P(X - \theta = -y \mid Y = y) = 0.5$$

More informally, we would expect about **half the observations to be above θ** and half below, with **both halves having the same distribution of Y** values.

Symmetry

Recall the **decomposition of Laplace/Double Exponential** random variables that were **symmetric about θ**:

$$X = SY + \theta$$

where S was the **sign of X**, Y was the **magnitude**.

Notice $Y = |X - \theta|$. If we **condition on Y**, then

$$P(X - \theta = y \mid Y = y) = P(X - \theta = -y \mid Y = y) = 0.5$$

More informally, we would expect about **half the observations to be above θ** and half below, with **both halves having the same distribution of Y** values.

Conditioning

On the previous slide we **conditioned** on the observed magntudes $y_i = x_i - \theta$ (for known median θ).

Conditioning

On the previous slide we **conditioned** on the observed magntudes $y_i = x_i - \theta$ (for known median θ).

Under H_0, we know that the statistic,

$$T(S_1, S_2, \ldots, S_n) = \sum_{i=1}^{n} S_i y_i = \sum_{i=1}^{n} S_i |x_i - \theta|$$

is **invariant to permutation** (i.e., we can swap around any S_i and S_j and T has the same distribution.

Conditioning

On the previous slide we **conditioned** on the observed magntudes $y_i = x_i - \theta$ (for known median θ).

Under H_0, we know that the statistic,

$$T(S_1, S_2, \ldots, S_n) = \sum_{i=1}^{n} S_i y_i = \sum_{i=1}^{n} S_i |x_i - \theta|$$

is **invariant to permutation** (i.e., we can swap around any S_i and S_j and T has the same distribution.

To find the **conditional distribution of T**: enumerate all 2^n possible $\{-1, 1\}^n$ (vectors of ± 1).

Conditioning

On the previous slide we **conditioned** on the observed magntudes $y_i = x_i - \theta$ (for known median θ).

Under H_0, we know that the statistic,

$$T(S_1, S_2, \ldots, S_n) = \sum_{i=1}^{n} S_i y_i = \sum_{i=1}^{n} S_i |x_i - \theta|$$

is **invariant to permutation** (i.e., we can swap around any S_i and S_j and T has the same distribution.

To find the **conditional distribution of T**: enumerate all 2^n possible $\{-1, 1\}^n$ (vectors of ± 1).

It will be difficult to enumerate all 2^n possible values, but we can use **Monte Carlo sampling**:

```
> testSymmetry <- function(x, theta, k = 1000) {
+     n <- length(x)
+     s_0 <- sum(x - theta)
+     y <- abs(x - theta)
+     dist <- map_dbl(rerun(k, 2 * rbinom(n, size = 1, p = 0.5) - 1),
+                 ~ sum(y * .x))
+     2 * min(dist >= s_0, dist <= s_0) / k
+ }
> testSymmetry(tracks10$duration_sec, 201)
```

```
[1] 0
```

Example: Two Sample Problems

Recall last time we considered the problem of estimating a **difference of means** for students from Phoenix and San Antonio.

Example: Two Sample Problems

Recall last time we considered the problem of estimating a **difference of means** for students from Phoenix and San Antonio.

We conceptualized this problem as **two samples**

$$X_1, \ldots, X_n \sim F, \quad \text{iid}, \quad \text{(Phonenix)}$$
$$Y_1, \ldots, Y_m \sim G, \quad \text{iid}, \quad \text{(San Antonio)}$$

Example: Two Sample Problems

Recall last time we considered the problem of estimating a **difference of means** for students from Phoenix and San Antonio.

We conceptualized this problem as **two samples**

$$X_1, \dots, X_n \sim F, \quad \text{iid}, \quad \text{(Phonenix)}$$
$$Y_1, \dots, Y_m \sim G, \quad \text{iid}, \quad \text{(San Antonio)}$$

Previously, we found a confidence interval for $\theta = E(X) - E(Y)$, but we could also ask a more general question:

$$H_0 : F = G \quad \text{vs} \quad H_1 : F \neq G$$

Combined Sample Notation

It's often convenient to think about a combined sample of the form:

$$(W_i, Z_i), \quad i = 1, \ldots, n + m$$

Where

$$W_i = \begin{cases} X_i & i \leq n \\ Y_i & i > n \end{cases}$$

and

$$Z_i = I(i \leq n)$$

(i.e., W_i is the data, Z_i is the label)

Combined Sample Notation

It's often convenient to think about a combined sample of the form:

$$(W_i, Z_i), \quad i = 1, \ldots, n+m$$

Where

$$W_i = \begin{cases} X_i & i \leq n \\ Y_i & i > n \end{cases}$$

and

$$Z_i = I(i \leq n)$$

(i.e., W_i is the data, Z_i is the label)

The permutation approach then just requires shuffling the Z_i values.

Picking a Test Statistic

Once we have a **null hypothesis** ($H_0 : F = G$), we need to **condition on the sample** select a **test statistic**

$$T(Z_1, \ldots, Z_{n+m})$$

Picking a Test Statistic

Once we have a **null hypothesis** ($H_0 : F = G$), we need to **condition on the sample** select a **test statistic**

$$T(Z_1, \ldots, Z_{n+m})$$

Notice that under the null hypothesis, **group labels are uninformative**, so we can permute Z:

$$T(Z_1, \ldots, Z_{n+m}) \sim T(\pi(Z_1, \ldots, Z_{n+m}))$$

Picking a Test Statistic

Once we have a **null hypothesis** ($H_0 : F = G$), we need to **condition on the sample** select a **test statistic**

$$T(Z_1, \ldots, Z_{n+m})$$

Notice that under the null hypothesis, **group labels are uninformative**, so we can permute Z:

$$T(Z_1, \ldots, Z_{n+m}) \sim T(\pi(Z_1, \ldots, Z_{n+m}))$$

A permutation test will permute the Z_i to get a conditional distribution for T.

Some Example Statistics

- Difference of means: $T(\mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} Z_i w_i - \frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)w_i$ (Welch's permutational t-test)

Some Example Statistics

- Difference of means: $T(\mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} Z_i w_i - \frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)w_i$ (Welch's permutational t-test)
- Difference of medians: $T(\mathbf{Z}) = \text{median}(w \mid Z = 1) - \text{median}(w \mid Z = 0)$

Some Example Statistics

- Difference of means: $T(\mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} Z_i w_i - \frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)w_i$ (Welch's permutational t-test)
- Difference of medians: $T(\mathbf{Z}) = \text{median}(w \mid Z = 1) - \text{median}(w \mid Z = 0)$
- Ratios of variance: $T(\mathbf{Z}) = \frac{S_1^2}{S_0^2}$ (sample variance for each group)

Some Example Statistics

- Difference of means: $T(\mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} Z_i w_i - \frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)w_i$ (Welch's permutational t-test)
- Difference of medians: $T(\mathbf{Z}) = \text{median}(w \mid Z = 1) - \text{median}(w \mid Z = 0)$
- Ratios of variance: $T(\mathbf{Z}) = \frac{S_1^2}{S_0^2}$ (sample variance for each group)
- Sum of scores: $T(\mathbf{Z}) = \sum_{i=1}^{n} Z_i g(w_i)$ (Wilcoxon-Mann-Whitney, Normal Scores)

Some Example Statistics

- Difference of means: $T(\mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} Z_i w_i - \frac{1}{n}\sum_{i=1}^{n}(1 - Z_i) w_i$ (Welch's permutational t-test)
- Difference of medians: $T(\mathbf{Z}) = \text{median}(w \mid Z = 1) - \text{median}(w \mid Z = 0)$
- Ratios of variance: $T(\mathbf{Z}) = \frac{S_1^2}{S_0^2}$ (sample variance for each group)
- Sum of scores: $T(\mathbf{Z}) = \sum_{i=1}^{n} Z_i g(w_i)$ (Wilcoxon-Mann-Whitney, Normal Scores)
- Comparisons of ECDFs (Kolmogorov-Smirnov, Andersong-Darling, Cramer-von Mises)

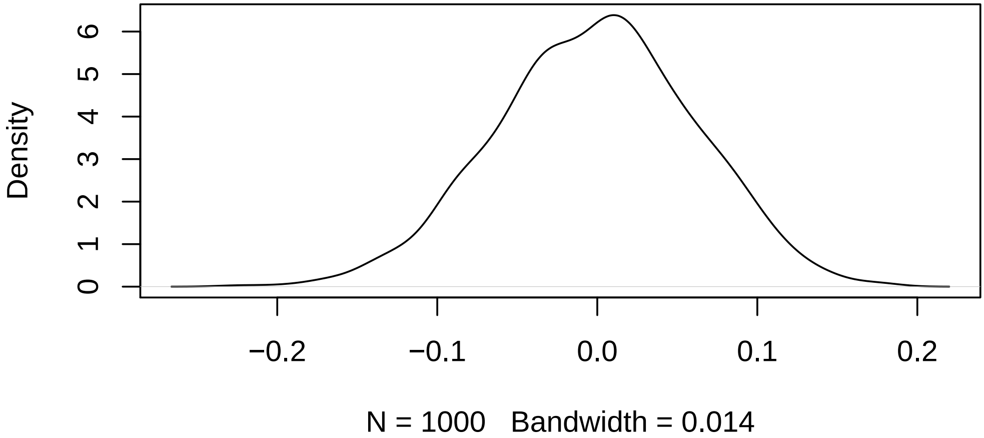
Phoenix and San Antonio

```
> mean_diff <- function(w, z) {mean(w[z], na.rm = TRUE) -
+                                   mean(w[!z], na.rm = TRUE)}
```

Phoenix and San Antonio

```
> mean_diff <- function(w, z) {mean(w[z], na.rm = TRUE) -
+                                 mean(w[!z], na.rm = TRUE)}

> n <- nrow(gamoran)
> dist.t <- replicate(1000, {
+     ## shuffle the "Z_i" values
+     permuted_label <- sample(gamoran$PH.AZ)
+     ## compute the test statistic
+     mean_diff(gamoran$READ_PCTZ, permuted_label)
+ })
```

Distribution under the null



Computing a p-value

We haven't specified an alternative yet, but let's consider

$$H_1 : F \neq G$$

such that the **distribution of T would be shifted** if H_1 is true.

Computing a p-value

We haven't specified an alternative yet, but let's consider

$$H_1 : F \neq G$$

such that the **distribution of T would be shifted** if H_1 is true.

In other words, we can use created "two tailed" p-value:

$$p - \text{value} \;= 2\min(P(T \leq t), P(T \geq t))$$

where t is the **observed value of the test statistic**.

Computing a p-value

We haven't specified an alternative yet, but let's consider

$$H_1 : F \neq G$$

such that the **distribution of T would be shifted** if H_1 is true.

In other words, we can use created "two tailed" p-value:

$$p - \text{value} \; = 2\min(P(T \leq t), P(T \geq t))$$

where t is the **observed value of the test statistic**.

```
> (t_observed <- mean_diff(gamoran$READ_PCTZ, gamoran$PH.AZ))
[1] 0.2568
> 2 * min(mean(dist.t <= t_observed), mean(dist.t >= t_observed))
[1] 0
```

Two Sample Permutation Test Framework

Like the bootstrap (or many other procedures we've seen), we can think about a general algorithm for permutation tests:

- Select a test statistic T that compares two samples

Two Sample Permutation Test Framework

Like the bootstrap (or many other procedures we've seen), we can think about a general algorithm for permutation tests:

- Select a test statistic T that compares two samples
- Compute the observed value $\hat{T}$

Two Sample Permutation Test Framework

Like the bootstrap (or many other procedures we've seen), we can think about a general algorithm for permutation tests:

- Select a test statistic T that compares two samples
- Compute the observed value $\hat{T}$
- Randomly generate B permutations of the $n + m$ group labels and compute T_b

Two Sample Permutation Test Framework

Like the bootstrap (or many other procedures we've seen), we can think about a general algorithm for permutation tests:

- Select a test statistic T that compares two samples
- Compute the observed value $\hat{T}$
- Randomly generate B permutations of the $n+m$ group labels and compute T_b
- Depending on the alternative, compute the p-value as

$$p^+ = \frac{1}{B}\sum_{b=1}^{B} I(T_b \geq \hat{T}), \quad p^- = \frac{1}{B}\sum_{b=1}^{B} I(T_b \leq \hat{T}), \quad p = 2 \times \min(p^+, p^-)$$

(Note: some sources add one to both numerator and denominator. For large B both approaches are about the same.)

Picking test statistics

By construction, permutation tests have **size no greater than level** (i.e., they will not reject more than $100 \times \alpha\%$ of the time).

Picking test statistics

By construction, permutation tests have **size no greater than level** (i.e., they will not reject more than $100 \times \alpha\%$ of the time).

But the power can be poor if we don't pick **good test statistics**.

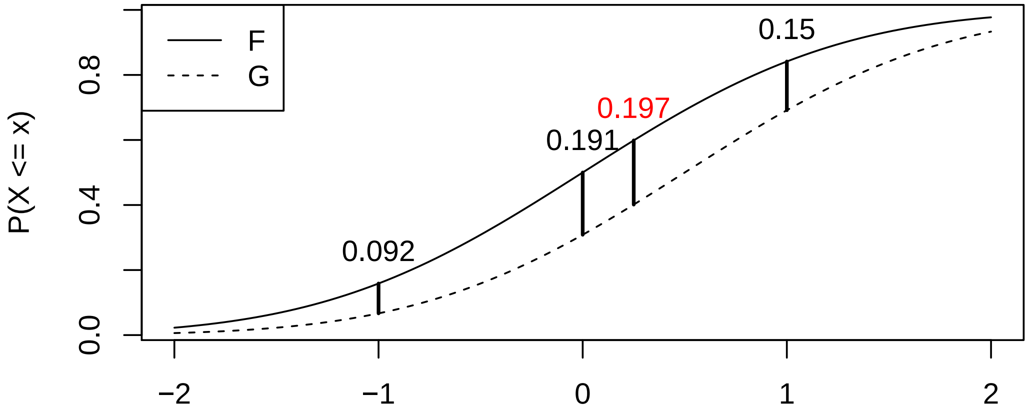
Picking test statistics

By construction, permutation tests have **size no greater than level** (i.e., they will not reject more than $100 \times \alpha\%$ of the time).

But the power can be poor if we don't pick **good test statistics**.

Next we'll look at a class of test statistics that have proved useful for many problems.

Kolmogorov-Smirnov



$$T(W,Z) = \max_{1 \leq i \leq n+m} \left| \widehat{F}_Z(W_i) - \widehat{G}_Z(W_i) \right|$$

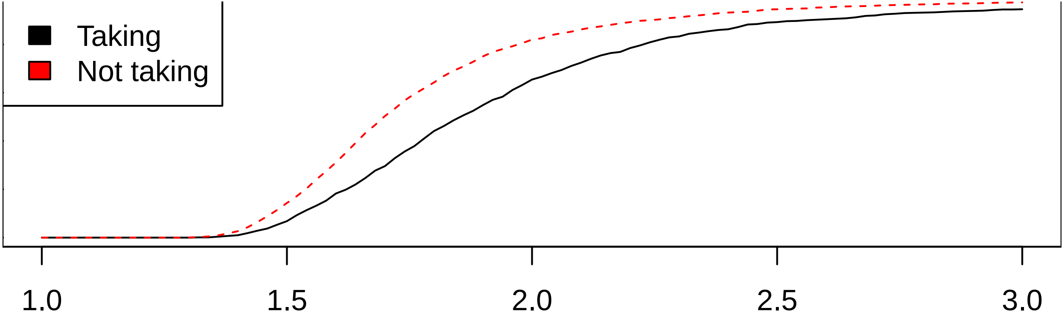
Blood pressure for subjects taking aspirin vs. not

Previously we looked an example from the NHANES study where respondents had a **blood pressure exam** and answered a survey question about **taking aspirin**.

Blood pressure for subjects taking aspirin vs. not

Previously we looked an example from the NHANES study where respondents had a **blood pressure exam** and answered a survey question about **taking aspirin**.

One way to summarize the BP measurements was to take the **ratio of systolic to diastolic** pressure:

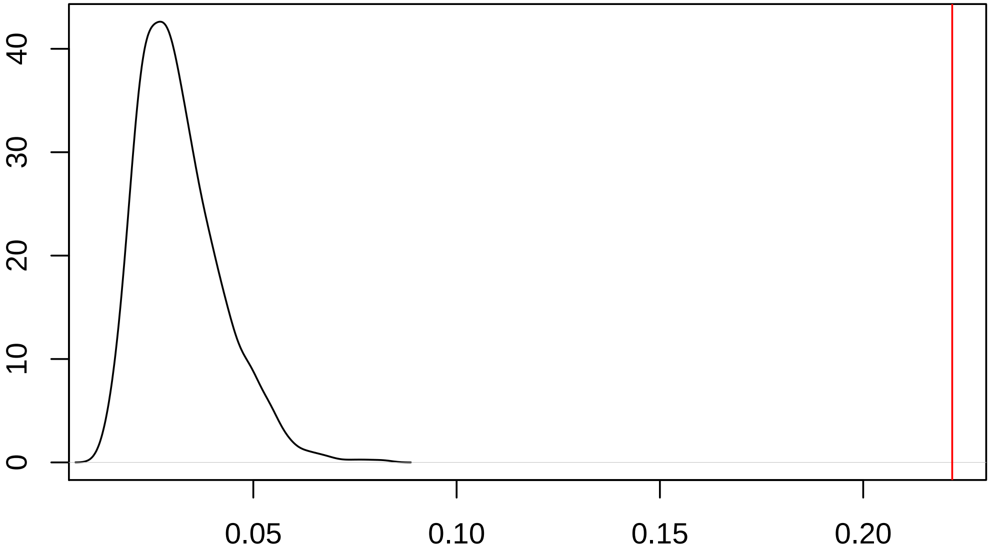


Implementing KS

```
> ks <- function(w, z) {
+     f <- ecdf(w[z == 1])
+     g <- ecdf(w[z == 0])
+     max(abs(f(w) - g(w)))
+ }
```

Implementing KS

```
> ks <- function(w, z) {
+   f <- ecdf(w[z == 1])
+   g <- ecdf(w[z == 0])
+   max(abs(f(w) - g(w)))
+ }

> perms <- replicate(1000, sample(nhanes$taking_aspirin))
> ts <- apply(perms, 2, function(zstar) {
+     ks(nhanes$ratio, zstar)
+ })
> observed_ks <- ks(nhanes$ratio, nhanes$taking_aspirin)
> (ksp <- 2 * min(mean(ts >= observed_ks), mean(ts <= observed_ks)))

[1] 0
```


0
10
20
30
40
0.05
0.10
0.15
0.20

Distribution Free Statistics

Interesting fact: when $H_0 : F = G$ is true, we can figure out the distribution of D **without even seeing the data** (W, Z), provided we know n and m.

Distribution Free Statistics

Interesting fact: when $H_0 : F = G$ is true, we can figure out the distribution of D **without even seeing the data** (W, Z), provided we know n and m.

This is because we can express the test statistic with respect to the **ranks** of the W_i. E.g.,

$$W = (3, 9, 2, 4) \rightarrow R = (2, 4, 1, 3)$$

Distribution Free Statistics

Interesting fact: when $H_0 : F = G$ is true, we can figure out the distribution of D **without even seeing the data (W, Z)**, provided we know n and m.

This is because we can express the test statistic with respect to the **ranks** of the W_i. E.g.,

$$W = (3, 9, 2, 4) \to R = (2, 4, 1, 3)$$

Claim: Let

$$D^+ = \max_i \hat{F}(W_i) - \hat{G}(W_i), \quad D^- = \max_i \hat{G}(W_i) - \hat{F}(W_i)$$

Then the statistic $\max(D^+, D^-)$ is **distribution free**.

Proof

For all $i = 1, \ldots, n+m$, let R_i be the **rank** of W_i in the combined sample. We'll show that D^+ only depends on the ranks, and the rest of the claim follows similarly.

Proof

For all $i = 1, \ldots, n+m$, let R_i be the **rank** of W_i in the combined sample. We'll show that D^+ only depends on the ranks, and the rest of the claim follows similarly.

Remembering that $Z_i = 1$ for $i \leq n$ and $Z_i = 0$ for $n \leq i \leq n+m$, notice that for any W_i,

$$\hat{F}(W_i) = \frac{1}{n} \sum_{j=1}^{n} I(W_j \leq W_i)$$

Proof

For all $i = 1, \ldots, n+m$, let R_i be the **rank** of W_i in the combined sample. We'll show that D^+ only depends on the ranks, and the rest of the claim follows similarly.

Remembering that $Z_i = 1$ for $i \leq n$ and $Z_i = 0$ for $n \leq i \leq n+m$, notice that for any W_i,

$$\begin{aligned}\hat{F}(W_i) &= \frac{1}{n}\sum_{j=1}^{n} I(W_j \leq W_i) \\ &= \frac{1}{n}\sum_{j=1}^{n+m} Z_j I(W_j \leq W_i)\end{aligned}$$

Proof

For all $i = 1, \ldots, n+m$, let R_i be the **rank** of W_i in the combined sample. We'll show that D^+ only depends on the ranks, and the rest of the claim follows similarly.

Remembering that $Z_i = 1$ for $i \leq n$ and $Z_i = 0$ for $n \leq i \leq n+m$, notice that for any W_i,

$$
\begin{aligned}
\hat{F}(W_i) &= \frac{1}{n} \sum_{j=1}^{n} I(W_j \leq W_i) \\
&= \frac{1}{n} \sum_{j=1}^{n+m} Z_j I(W_j \leq W_i) \\
&= \frac{1}{n} \sum_{j=1}^{n+m} Z_j I(R_j \leq R_i)
\end{aligned}
$$

Proof, cont.: Replace ranks with integers

So few have $\hat{F}(W_i) = (1/n)\sum_{j=1}^{n+m} Z_j I(R_j \leq R_i)$ and, likewise, $\hat{G}(W_i) = (1/m)\sum_{j=1}^{n+m}(1 - Z_j) I(R_j \leq R_i)$.

Proof, cont.: Replace ranks with integers

So few have $\hat{F}(W_i) = (1/n)\sum_{j=1}^{n+m} Z_j I(R_j \leq R_i)$ and, likewise, $\hat{G}(W_i) = (1/m)\sum_{j=1}^{n+m}(1 - Z_j)I(R_j \leq R_i)$.

Since the ranks are the integers $1, \ldots, n+m$, we can write D^+ as

$$D^+ = \max_{1 \leq i \leq n+m} \hat{F}(W_i) - \hat{G}(W_i)$$

Proof, cont.: Replace ranks with integers

So few have $\hat{F}(W_i) = (1/n)\sum_{j=1}^{n+m} Z_j I(R_j \le R_i)$ and, likewise, $\hat{G}(W_i) = (1/m)\sum_{j=1}^{n+m}(1 - Z_j)I(R_j \le R_i)$.

Since the ranks are the integers $1, \dots, n+m$, we can write D^+ as

$$\begin{aligned} D^+ &= \max_{1 \le i \le n+m} \hat{F}(W_i) - \hat{G}(W_i) \\ &= \max_{1 \le i \le n+m} \frac{1}{n} \sum_{j=1}^{n+m} Z_j I(R_j \le R_i) - \sum_{j=1}^{n+m} (1 - Z_j) I(R_j \le R_i) \end{aligned}$$

Proof, cont.: Replace ranks with integers

So few have $\hat{F}(W_i) = (1/n)\sum_{j=1}^{n+m} Z_j I(R_j \leq R_i)$ and, likewise, $\hat{G}(W_i) = (1/m)\sum_{j=1}^{n+m}(1 - Z_j)I(R_j \leq R_i)$.

Since the ranks are the integers $1, \ldots, n+m$, we can write D^+ as

$$\begin{aligned}
D^+ &= \max_{1\leq i\leq n+m} \hat{F}(W_i) - \hat{G}(W_i) \\
&= \max_{1\leq i\leq n+m} \frac{1}{n}\sum_{j=1}^{n+m} Z_j I(R_j \leq R_i) - \sum_{j=1}^{n+m}(1 - Z_j)I(R_j \leq R_i) \\
&= \max_{1\leq i\leq n+m} \frac{1}{n}\sum_{j=1}^{n+m} Z_j I(R_j \leq i) - \frac{1}{m}\sum_{j=1}^{n+m}(1 - Z_j)I(R_j \leq i)
\end{aligned}$$

Distribution free tests

In the previous slide we replaced the **sample values** with **ranks**, which must be composed of $1, \dots, n + m$.

Distribution free tests

In the previous slide we replaced the **sample values** with **ranks**, which must be composed of $1, \ldots, n + m$.

Implication: For any two sample problem of n and m, the **distribution of D^+** is **always the same** (i.e., it does not depend on the sample values).

Distribution free tests

In the previous slide we replaced the **sample values** with **ranks**, which must be composed of $1, \ldots, n + m$.

Implication: For any two sample problem of n and m, the **distribution of D^+** is **always the same** (i.e., it does not depend on the sample values).

More generally, so we can **compute the distribution of any statistic $T(R, Z)$ before we see any data**. (provided T doesn't depend on w otherwise.)

Distribution free tests

In the previous slide we replaced the **sample values** with **ranks**, which must be composed of $1, \ldots, n+m$.

Implication: For any two sample problem of n and m, the **distribution of D^+** is **always the same** (i.e., it does not depend on the sample values).

More generally, so we can **compute the distribution of any statistic $T(R, Z)$ before we see any data**. (provided T doesn't depend on w otherwise.)

This is precisely what we call **distribution free**. (Technically, we have parameters n and m, but nothing that depends on the data.)

Other Distribution Free Tests

Many times we can create distribution tests by applying an existing test statistic to the ranks of the data.

Other Distribution Free Tests

Many times we can create distribution tests by applying an existing test statistic to the ranks of the data. For example, we already used **the difference of means statistic**

$$T(Z,W) = \frac{1}{n}\sum_{i=1}^{n+m} Z_i W_i - \frac{1}{m}\sum_{i=1}^{n+m}(1-Z_i)W_i$$

Many times we can create distribution tests by applying an existing test statistic to the ranks of the data. For example, we already used **the difference of means statistic**

$$T(Z, W) = \frac{1}{n}\sum_{i=1}^{n+m} Z_i W_i - \frac{1}{m}\sum_{i=1}^{n+m}(1 - Z_i)W_i$$

Instead, we could do a **difference of average ranks**

$$T(Z, R) = \frac{1}{n}\sum_{i=1}^{n+m} Z_i R_i - \frac{1}{m}\sum_{i=1}^{n+m}(1 - Z_i)R_i$$

Other Distribution Free Tests

Many times we can create distribution tests by applying an existing test statistic to the ranks of the data. For example, we already used **the difference of means statistic**

$$T(Z, W) = \frac{1}{n} \sum_{i=1}^{n+m} Z_i W_i - \frac{1}{m} \sum_{i=1}^{n+m} (1 - Z_i) W_i$$

Instead, we could do a **difference of average ranks**

$$T(Z, R) = \frac{1}{n} \sum_{i=1}^{n+m} Z_i R_i - \frac{1}{m} \sum_{i=1}^{n+m} (1 - Z_i) R_i$$

As with the KS test, it doesn't matter what the actual W values are, we can get the distribution of $T(Z, R)$ **without observing any data**.

Other Distribution Free Tests

Many times we can create distribution tests by applying an existing test statistic to the ranks of the data. For example, we already used **the difference of means statistic**

$$T(Z,W) = \frac{1}{n}\sum_{i=1}^{n+m} Z_i W_i - \frac{1}{m}\sum_{i=1}^{n+m}(1-Z_i)W_i$$

Instead, we could do a **difference of average ranks**

$$T(Z,R) = \frac{1}{n}\sum_{i=1}^{n+m} Z_i R_i - \frac{1}{m}\sum_{i=1}^{n+m}(1-Z_i)R_i$$

As with the KS test, it doesn't matter what the actual W values are, we can get the distribution of $T(Z,R)$ **without observing any data**.

This test is known as the **Wilcoxon-Mann-Whitney** test.

- Kolmogorov-Smirnov: `ks.test`
- Wilcoxon-Mann-Whitney: `wilcox.test`
- Normal Scores: for $H_i = \Phi^{-1}(R_i/(n+m+1))$:

$$T(Z, H) = \frac{1}{n}\sum_{i=1}^{n+m} Z_i H_i - \frac{1}{m}\sum_{i=1}^{n+m}(1 - Z_i)H_i$$

  is implemented in the `SuppDist` package.

And we can always estimate any other test statistic distribution using a Monte Carlo approach.

```
> with(nhanes, ## creates variables ratio, taking_apsirin
+      ks.test(x = ratio[taking_aspirin],
+              y = ratio[!taking_aspirin]))

        Two-sample Kolmogorov-Smirnov test

data:  ratio[taking_aspirin] and ratio[!taking_aspirin]
D = 0.22, p-value <2e-16
alternative hypothesis: two-sided
```

```
> with(nhanes,
+       wilcox.test(x = ratio[taking_aspirin],
+                   y = ratio[!taking_aspirin]))

        Wilcoxon rank sum test with continuity correction

data:  ratio[taking_aspirin] and ratio[!taking_aspirin]
W = 1645455, p-value <2e-16
alternative hypothesis: true location shift is not equal to 0
```

Summary

Permutation tests are a large class of hypothesis tests based on **permutation invariant test statistics**: by conditioning on some aspect of the sample, the test statistic has **the same distribution under any permutation of the remaining random data**.

Summary

Permutation tests are a large class of hypothesis tests based on **permutation invariant test statistics**: by conditioning on some aspect of the sample, the test statistic has **the same distribution under any permutation of the remaining random data**.

In general, something of a an art to find what can be conditioned and what remains random.

Summary

Permutation tests are a large class of hypothesis tests based on **permutation invariant test statistics**: by conditioning on some aspect of the sample, the test statistic has **the same distribution under any permutation of the remaining random data**.

In general, something of a an art to find what can be conditioned and what remains random.

Very common setting is the **two sample problem** where we are testing that groups have the same distribution: $H_0 : F = G$.

Summary

Permutation tests are a large class of hypothesis tests based on **permutation invariant test statistics**: by conditioning on some aspect of the sample, the test statistic has **the same distribution under any permutation of the remaining random data**.

In general, something of a an art to find what can be conditioned and what remains random.

Very common setting is the **two sample problem** where we are testing that groups have the same distribution: $H_0 : F = G$.

Distribution free tests replace data with ranks (or similar) to make test statistics not depend on the underlying distribution of the data.