Week 03: Monte Carlo Hypothesis Testing

Mark M. Fredrickson (mfredric@umich.edu)

Computational Methods in Statistics and Data Science (Stats 406)

Review

- Integration as a key concept in statistics

Review

- Integration as a key concept in statistics
- Approximating integrals by Monte Carlo sampling:

Review

- Integration as a key concept in statistics
- Approximating integrals by Monte Carlo sampling:
    - For integral $\int_a^b g(x)\,dx$ identify an RV with support $[a, b]$.

Review

- Integration as a key concept in statistics
- Approximating integrals by Monte Carlo sampling:
  - For integral $\int_a^b g(x)\,dx$ identify an RV with support $[a, b]$.
  - Transform problem into

$$\int_a^b \frac{g(x)}{f(x)} f(x)\,dx = \mathrm{E}\left(\frac{g(X)}{f(X)}\right)$$

Review

- Integration as a key concept in statistics
- Approximating integrals by Monte Carlo sampling:
  - For integral $\int_a^b g(x)\,dx$ identify an RV with support $[a, b]$.
  - Transform problem into

$$\int_a^b \frac{g(x)}{f(x)} f(x)\,dx = \mathrm{E}\left(\frac{g(X)}{f(X)}\right)$$

  - Sample from X and compute

$$\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i)}{f(X_i)}$$

Review

- Integration as a key concept in statistics
- Approximating integrals by Monte Carlo sampling:
  - For integral $\int_a^b g(x)\,dx$ identify an RV with support $[a, b]$.
  - Transform problem into
    $$\int_a^b \frac{g(x)}{f(x)} f(x)\,dx = \mathrm{E}\left(\frac{g(X)}{f(X)}\right)$$
  - Sample from X and compute
    $$\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i)}{f(X_i)}$$
- Estimates converge, CLT implies intervals

Review

- Integration as a key concept in statistics
- Approximating integrals by Monte Carlo sampling:
    - For integral $\int_a^b g(x)\,dx$ identify an RV with support $[a, b]$.
    - Transform problem into

$$\int_a^b \frac{g(x)}{f(x)} f(x)\,dx = \mathrm{E}\left(\frac{g(X)}{f(X)}\right)$$

    - Sample from X and compute

$$\frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)}$$

- Estimates converge, CLT implies intervals
- Probability as expectation of indicator functions

Monte Carlo Hypothesis Testing

Hypothesis Tests

We'll begin our exploration of **operating characteristics** of statistical procedures with **hypothesis tests**.

A hypothesis test requires stating a **null hypothesis** H_0 and an **alternative hypothesis** H_1. Some examples:

$$H_0 : X \overset{\text{iid}}{\sim} F_0 \qquad \text{vs. } H_1 : X \overset{\text{iid}}{\sim} F_1$$

Hypothesis Tests

We'll begin our exploration of **operating characteristics** of statistical procedures with **hypothesis tests**.

A hypothesis test requires stating a **null hypothesis** H_0 and an **alternative hypothesis** H_1. Some examples:

$$
\begin{aligned}
H_0 &: X \overset{\text{iid}}{\sim} F_0 &\quad \text{vs. } H_1 : X \overset{\text{iid}}{\sim} F_1 \\
H_0 &: E(X) \leq \mu_0 &\quad \text{vs. } H_1 : E(X) > \mu_0
\end{aligned}
$$

Hypothesis Tests

We'll begin our exploration of **operating characteristics** of statistical procedures with **hypothesis tests**.

A hypothesis test requires stating a **null hypothesis** H_0 and an **alternative hypothesis** H_1. Some examples:

$$
\begin{aligned}
H_0 &: X \overset{\text{iid}}{\sim} F_0 && \text{vs. } H_1 : X \overset{\text{iid}}{\sim} F_1 \\
H_0 &: E(X) \leq \mu_0 && \text{vs. } H_1 : E(X) > \mu_0 \\
H_0 &: F(x, y) = F_X(x)F_Y(y) && \text{v.s. } H_1 : F(x, y) \neq F_X(x)F_Y(y)
\end{aligned}
$$

We'll begin our exploration of **operating characteristics** of statistical procedures with **hypothesis tests**.

A hypothesis test requires stating a **null hypothesis** H_0 and an **alternative hypothesis** H_1. Some examples:

$$
\begin{aligned}
H_0 &: X \overset{\text{iid}}{\sim} F_0 && \text{vs. } H_1 : X \overset{\text{iid}}{\sim} F_1 \\
H_0 &: E(X) \le \mu_0 && \text{vs. } H_1 : E(X) > \mu_0 \\
H_0 &: F(x, y) = F_X(x)F_Y(y) && \text{v.s. } H_1 : F(x, y) \ne F_X(x)F_Y(y)
\end{aligned}
$$

Goal: Either **accept** the null hypothesis or **reject** the null hypothesis in favor of the alternative.

Type I and Type II Error

If we **reject a true null hypothesis**, we have committed a **Type I error**.

Type I and Type II Error

If we **reject a true null hypothesis**, we have committed a **Type I error**.

If we **accept a false null hypothesis**, we have committed a **Type II error**.

Type I and Type II Error

If we **reject a true null hypothesis**, we have committed a **Type I error**.

If we **accept a false null hypothesis**, we have committed a **Type II error**.

The probability of making a Type I error is the **size** of the test:

$$P(\text{Reject } H_0 \mid H_0)$$

Type I and Type II Error

If we **reject a true null hypothesis**, we have committed a **Type I error**.

If we **accept a false null hypothesis**, we have committed a **Type II error**.

The probability of making a Type I error is the **size** of the test:

$$P(\text{Reject } H_0 \mid H_0)$$

We define the probability of **not making a Type II error** when H_1 is true as the **power** of the test:

$$P(\text{Reject } H_0 \mid H_1)$$

Type I and Type II Error

If we **reject a true null hypothesis**, we have committed a **Type I error**.

If we **accept a false null hypothesis**, we have committed a **Type II error**.

The probability of making a Type I error is the **size** of the test:

$$P(\text{Reject } H_0 \mid H_0)$$

We define the probability of **not making a Type II error** when H_1 is true as the **power** of the test:

$$P(\text{Reject } H_0 \mid H_1)$$

Useful framework: pick a **maximum Type I error** α and then pick a test that has **good power**.

Test Statistics

In the previous slide we defined size and power using the **probability of rejecting** the null hypothesis (when it was true or false, respectively).

Test Statistics

In the previous slide we defined size and power using the **probability of rejecting** the null hypothesis (when it was true or false, respectively).

This probability comes from the **test statistic** we use to make our decision:

$$T(X_1, \ldots, X_n)$$

Test Statistics

In the previous slide we defined size and power using the **probability of rejecting** the null hypothesis (when it was true or false, respectively).

This probability comes from the **test statistic** we use to make our decision:

$$T(X_1, \ldots, X_n) = T$$

Test Statistics

In the previous slide we defined size and power using the **probability of rejecting** the null hypothesis (when it was true or false, respectively).

This probability comes from the **test statistic** we use to make our decision:

$$T(X_1, \ldots, X_n) = T$$

and a **rejection region** $\mathcal{R}$ such that

$$T \in \mathcal{R} \iff \text{reject the null hypothesis}$$

Test Statistics

In the previous slide we defined size and power using the **probability of rejecting** the null hypothesis (when it was true or false, respectively).

This probability comes from the **test statistic** we use to make our decision:

$$T(X_1, \ldots, X_n) = T$$

and a **rejection region** $\mathcal{R}$ such that

$$T \in \mathcal{R} \iff \text{reject the null hypothesis}$$

Where does $\mathcal{R}$ come from? How do we pick it?

Distributions for Test Statistics

Recall this key feature of statistics: because $X_1, \ldots, X_n$ are random, **test statistics are also random**!

Distributions for Test Statistics

Recall this key feature of statistics: because $X_1, \ldots, X_n$ are random, **test statistics are also random**!

The **distribution of T is determined by the hypothesis**:

- If H_0 is true, T is governed by the **null distribution**.
- If H_1 is true, T is governed by the **alternative distribution**.

Distributions for Test Statistics

Recall this key feature of statistics: because $X_1, \ldots, X_n$ are random, **test statistics are also random**!

The **distribution of T is determined by the hypothesis**:

- If H_0 is true, T is governed by the **null distribution**.
- If H_1 is true, T is governed by the **alternative distribution**.

To determine what these distributions are, there are typically two methods:

- Analyze $T(X_1, \ldots, X_n)$ under the assumptions of H_0 and H_1 (classical)
- Use Monte Carlo techniques to draw samples of $(X_1, \ldots, X_n)$ and estimate distributions for T

Example: Testing $\mu_0 = 0$ vs. $\mu_0 = 1$

Suppose we assume that

$$X_1, X_2 \overset{\text{iid}}{\sim} N(\mu, 1) \quad (\mu \text{ unknown})$$

and want to test:

$$H_0 : \mu = 0 \quad \text{vs. } H_1 : \mu = 1$$

Example: Testing $\mu_0 = 0$ vs. $\mu_0 = 1$

Suppose we assume that

$$X_1, X_2 \overset{\text{iid}}{\sim} N(\mu, 1) \quad (\mu \text{ unknown})$$

and want to test:

$$H_0 : \mu = 0 \quad \text{vs. } H_1 : \mu = 1$$

We already know that $\bar{X}_n$ is the MLE for μ, perhaps that would be a good statistic:

- When H_0 is true,

$$\bar{X}_n \sim N\left(0, \frac{1}{2}\right) \quad \text{(null distribution)}$$

Example: Testing $\mu_0 = 0$ vs. $\mu_0 = 1$

Suppose we assume that

$$X_1, X_2 \overset{\text{iid}}{\sim} N(\mu, 1) \quad (\mu \text{ unknown})$$

and want to test:

$$H_0 : \mu = 0 \qquad \text{vs. } H_1 : \mu = 1$$

We already know that $\bar{X}_n$ is the MLE for μ, perhaps that would be a good statistic:

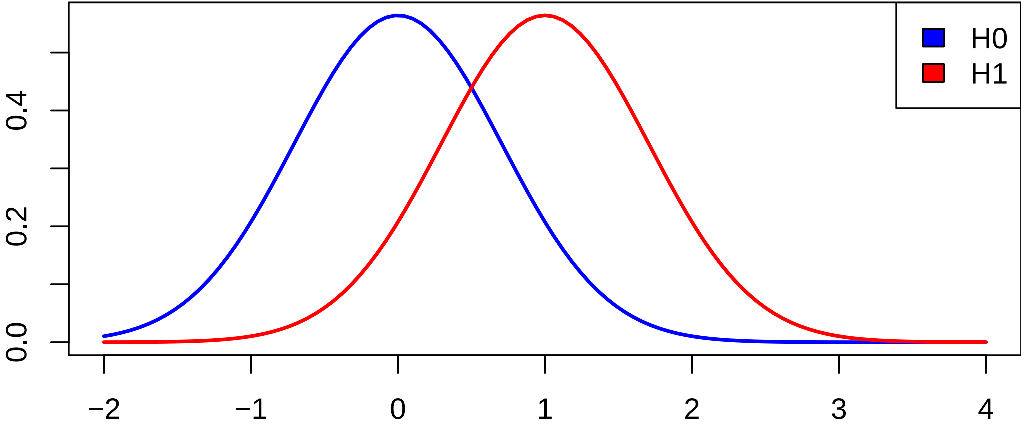
- When H_0 is true,

$$\bar{X}_n \sim N\left(0, \frac{1}{2}\right) \quad \text{(null distribution)}$$

- When H_1 is true,

$$\bar{X}_n \sim N\left(1, \frac{1}{2}\right) \quad \text{(alternative distribution)}$$

Test Statistic Distribution



Computing probability of Type I error and power

Now that we know $T \mid H_0$ and $T \mid H_1$ how should we pick $\mathcal{R}$?

Computing probability of Type I error and power

Now that we know $T \mid H_0$ and $T \mid H_1$ how should we pick $\mathcal{R}$?

Suppose we limit ourselves to just pick from the intervals $[a, b]$ (letting $a \to -\infty$ or $b \to \infty$), then the **probability of Type I error** is:

$$P(T \in \mathcal{R} \mid H_0) = P(T \in [a, b] \mid H_0)$$

Computing probability of Type I error and power

Now that we know $T \mid H_0$ and $T \mid H_1$ how should we pick $\mathcal{R}$?

Suppose we limit ourselves to just pick from the intervals $[a, b]$ (letting $a \to -\infty$ or $b \to \infty$), then the **probability of Type I error** is:

$$P(T \in \mathcal{R} \mid H_0) = P(T \in [a, b] \mid H_0) = P(T \leq b \mid H_0) - P(T \leq a \mid H_0)$$

Computing probability of Type I error and power

Now that we know $T \mid H_0$ and $T \mid H_1$ how should we pick $\mathcal{R}$?

Suppose we limit ourselves to just pick from the intervals $[a, b]$ (letting $a \to -\infty$ or $b \to \infty$), then the **probability of Type I error** is:

$$P(T \in \mathcal{R} \mid H_0) = P(T \in [a, b] \mid H_0) = P(T \leq b \mid H_0) - P(T \leq a \mid H_0)$$

Similarly the **power of test** is

$$P(T \in \mathcal{R} \mid H_1) = P(T \in [a, b] \mid H_1) = P(T \leq b \mid H_1) - P(T \leq a \mid H_1)$$

Computing probability of Type I error and power

Now that we know $T \mid H_0$ and $T \mid H_1$ how should we pick $\mathcal{R}$?

Suppose we limit ourselves to just pick from the intervals $[a, b]$ (letting $a \to -\infty$ or $b \to \infty$), then the **probability of Type I error** is:

$$P(T \in \mathcal{R} \mid H_0) = P(T \in [a, b] \mid H_0) = P(T \leq b \mid H_0) - P(T \leq a \mid H_0)$$

Similarly the **power of test** is

$$P(T \in \mathcal{R} \mid H_1) = P(T \in [a, b] \mid H_1) = P(T \leq b \mid H_1) - P(T \leq a \mid H_1)$$

Note: We are will use intervals for simplicity, but the basic ideas apply to more general $\mathcal{R}$.

Example: size and power for $\mu = 0$ vs. $\mu = 1$

Suppose that we use $\mathcal{R} = [0, 0.5]$ for the previously discussed hypothesis test $H_0 : \bar{X} \sim N(0, 1/n)$ vs $H_1 : \bar{X} \sim N(1, 1/n)$.

Example: size and power for $\mu = 0$ vs. $\mu = 1$

Suppose that we use $\mathcal{R} = [0, 0.5]$ for the previously discussed hypothesis test $H_0 : \bar{X} \sim N(0, 1/n)$ vs $H_1 : \bar{X} \sim N(1, 1/n)$.

First, size (under the null):

```
> n <- 2; a <- 0; b <- 0.5
> pnorm(b, sd = 1/sqrt(n)) - pnorm(a, sd = 1/sqrt(n))

[1] 0.2602
```

Example: size and power for $\mu = 0$ vs. $\mu = 1$

Suppose that we use $\mathcal{R} = [0, 0.5]$ for the previously discussed hypothesis test $H_0 : \bar{X} \sim N(0, 1/n)$ vs $H_1 : \bar{X} \sim N(1, 1/n)$.

First, size (under the null):

```
> n <- 2; a <- 0; b <- 0.5
> pnorm(b, sd = 1/sqrt(n)) - pnorm(a, sd = 1/sqrt(n))

[1] 0.2602
```

and power (under the alternative):

```
> pnorm(b, mean = 1, sd = 1/sqrt(n)) - pnorm(a, mean = 1, sd = 1/sqrt(n))

[1] 0.1611
```

Limiting Type I error

Since **we pick** $[a, b]$ (or $\mathcal{R}$ more generally), we can limit to $[a, b]$ with the property:

$$P(T \in [a, b] \mid H_0) \leq \alpha$$

We call α the **level** and $P(T \in [a, b] \mid H_0)$ the **size of the test** (i.e., we ensure size is no more than level).

Limiting Type I error

Since **we pick** $[a, b]$ (or $\mathcal{R}$ more generally), we can limit to $[a, b]$ with the property:

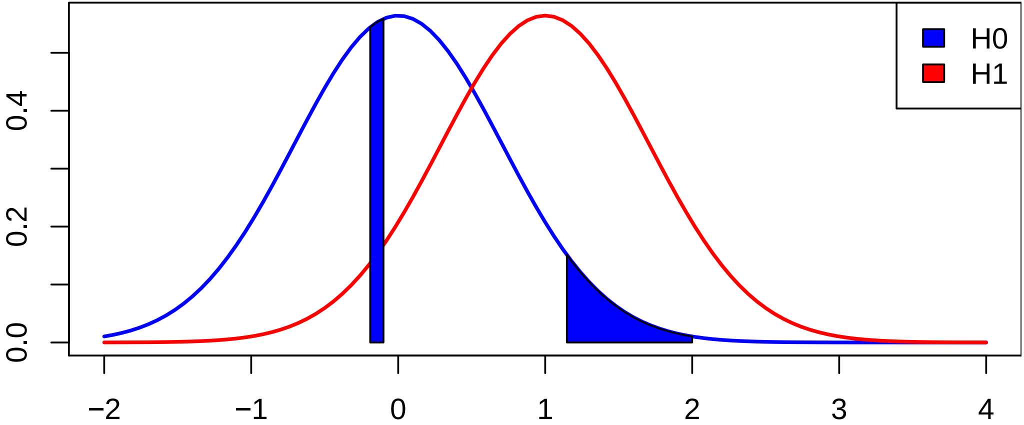
$$P(T \in [a, b] \mid H_0) \leq \alpha$$

We call α the **level** and $P(T \in [a, b] \mid H_0)$ the **size of the test** (i.e., we ensure size is no more than level).

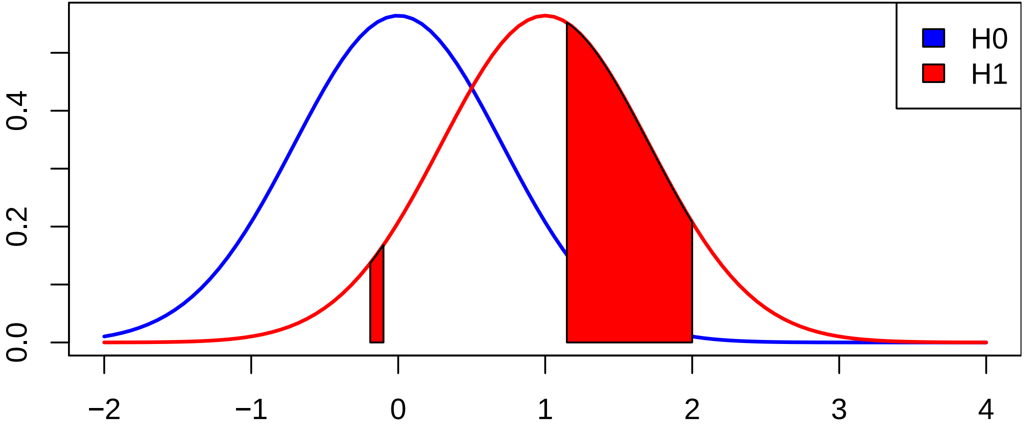
But this is not (usually) unique! Under our null hypothesis ($T \sim N(0, 1/2)$):

```
> n <- 2
> pnorm(-0.1, sd = 1/sqrt(n)) - pnorm(-0.1905725, sd = 1/sqrt(2))
[1] 0.05
> pnorm(2, sd = 1/sqrt(n)) - pnorm(1.147342, sd = 1/sqrt(2))
[1] 0.05
```

Rejection Regions: Size



Rejection Regions: Power



Computing rejection region and power ($n = 2$)

If we continued to increase b (keeping α fixed), we would eventually find the **one-tailed test** that has the highest possible power:

$$\mathcal{R} = [a, \infty)$$

Computing rejection region and power ($n = 2$)

If we continued to increase b (keeping α fixed), we would eventually find the **one-tailed test** that has the highest possible power:

$$\mathcal{R} = [a, \infty)$$

Computing the rejection region when $H_0 : \mu = 0$:

```
> (cutoff <- qnorm(0.95, mean = 0, sd = sqrt(1/sqrt(2))))
```

```
[1] 1.383
```

Computing rejection region and power ($n = 2$)

If we continued to increase b (keeping α fixed), we would eventually find the **one-tailed test** that has the highest possible power:

$$\mathcal{R} = [a, \infty)$$

Computing the rejection region when $H_0 : \mu = 0$:

```
> (cutoff <- qnorm(0.95, mean = 0, sd = sqrt(1/sqrt(2))))
```

```
[1] 1.383
```

Computing the power of the test when $H_1 : \mu = 1$:

```
> 1 - pnorm(cutoff, mean = 1, sd = sqrt(1/2))
```

```
[1] 0.294
```

Power Curves

A **power curve** shows how the power of a test changes with respect to another variable (sample size, tuning parameter, alternative hypothesis).

```
> sample_sizes <- 2:20
> power_n <- map_dbl(sample_sizes, function(x) {
+     cutoff <- qnorm(0.95, mean = 0, sd = sqrt(1 / x))
+     1 - pnorm(cutoff, mean = 1, sd = sqrt(1 / x))
+ })
```

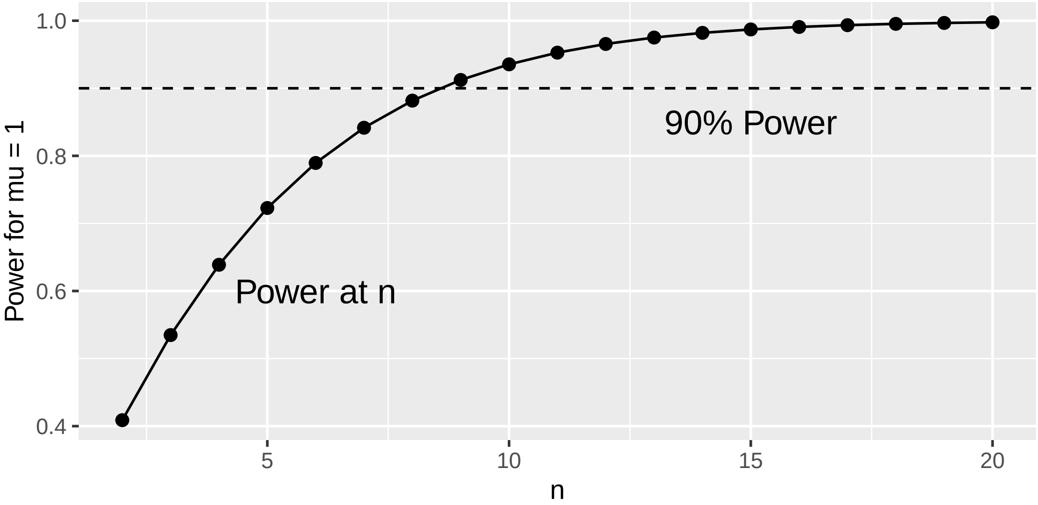
Power Curves

A **power curve** shows how the power of a test changes with respect to another variable (sample size, tuning parameter, alternative hypothesis).

```
> sample_sizes <- 2:20
> power_n <- map_dbl(sample_sizes, function(x) {
+     cutoff <- qnorm(0.95, mean = 0, sd = sqrt(1 / x))
+     1 - pnorm(cutoff, mean = 1, sd = sqrt(1 / x))
+ })

> pc <- ggplot(data.frame(n = sample_sizes, y = power_n),
+              aes(x = n, y = y))  +
+       geom_point(size = 2) + geom_line()
>
```

Power Curve Plot



Simple vs. Simple Tests

The "simple vs. simple" hypothesis test compares two distributions F_0 and F_1:

- A **null hypothesis**: : $(X_1, \ldots, X_n) \sim F_0$

Simple vs. Simple Tests

The "simple vs. simple" hypothesis test compares two distributions F_0 and F_1:

- A **null hypothesis**: : $(X_1, \ldots, X_n) \sim F_0$
- A **alternative hypothesis** $(X_1, \ldots, X_n) \sim F_1$.

Simple vs. Simple Tests

The "simple vs. simple" hypothesis test compares two distributions F_0 and F_1:

- A **null hypothesis**: : $(X_1, \ldots, X_n) \sim F_0$
- A **alternative hypothesis** $(X_1, \ldots, X_n) \sim F_1$.
- A **test statistic** $T(X_1, \ldots, X_n) : (X_1, \ldots, X_n) \to \mathbb{R}$.

Simple vs. Simple Tests

The "simple vs. simple" hypothesis test compares two distributions F_0 and F_1:

- A **null hypothesis**: : $(X_1, \dots, X_n) \sim F_0$
- A **alternative hypothesis** $(X_1, \dots, X_n) \sim F_1$.
- A **test statistic** $T(X_1, \dots, X_n) : (X_1, \dots, X_n) \to \mathbb{R}$.
- The **distribution of T** when F_0 holds (**null distribution**)

Simple vs. Simple Tests

The "simple vs. simple" hypothesis test compares two distributions F_0 and F_1:

- A **null hypothesis**: : $(X_1, \ldots, X_n) \sim F_0$
- A **alternative hypothesis** $(X_1, \ldots, X_n) \sim F_1$.
- A **test statistic** $T(X_1, \ldots, X_n) : (X_1, \ldots, X_n) \to \mathbb{R}$.
- The **distribution of T** when F_0 holds (**null distribution**)
- An α-level ($\alpha \in (0, 1)$) that specifies our **willingness to make an error** about the **null hypothesis**.

Simple vs. Simple Tests

The "simple vs. simple" hypothesis test compares two distributions F_0 and F_1:

- A **null hypothesis**: : $(X_1, \ldots, X_n) \sim F_0$
- A **alternative hypothesis** $(X_1, \ldots, X_n) \sim F_1$.
- A **test statistic** $T(X_1, \ldots, X_n) : (X_1, \ldots, X_n) \to \mathbb{R}$.
- The **distribution of T** when F_0 holds (**null distribution**)
- An α-level ($\alpha \in (0, 1)$) that specifies our **willingness to make an error** about the **null hypothesis**.
- **Rejection region $\mathcal{R}$**: reject H_0 when $T \in \mathcal{R}$ where $\Pr(T \in \mathcal{R} | H_0) \leq \alpha$.

Simple vs. Simple Tests

The "simple vs. simple" hypothesis test compares two distributions F_0 and F_1:

- A **null hypothesis**: : $(X_1, \ldots, X_n) \sim F_0$
- A **alternative hypothesis** $(X_1, \ldots, X_n) \sim F_1$.
- A **test statistic** $T(X_1, \ldots, X_n) : (X_1, \ldots, X_n) \to \mathbb{R}$.
- The **distribution of T** when F_0 holds (**null distribution**)
- An α-level ($\alpha \in (0, 1)$) that specifies our **willingness to make an error** about the **null hypothesis**.
- **Rejection region $\mathcal{R}$**: reject H_0 when $T \in \mathcal{R}$ where $\Pr(T \in \mathcal{R} | H_0) \leq \alpha$.
- Good tests will have high $\Pr(T \in \mathcal{R} | H_1)$ (**power**).

MC for Simple v. Simple

Observe that Type I error and power can be written as **expectations**:

$$P(T \in \mathcal{R} \mid H_0) = E(I(T \in \mathcal{R}) \mid H_0), \quad P(T \in \mathcal{R} \mid H_1) = E(I(T \in \mathcal{R}) \mid H_1)$$

MC for Simple v. Simple

Observe that Type I error and power can be written as **expectations**:

$$P(T \in \mathcal{R} \mid H_0) = E(I(T \in \mathcal{R}) \mid H_0), \quad P(T \in \mathcal{R} \mid H_1) = E(I(T \in \mathcal{R}) \mid H_1)$$

Conditioning on H_0 being true means that $X_i \sim F_0$ (likewise for H_1):

- Generate k **samples of size** n **from** F_0: $X_{j1}, X_{j2}, \ldots, X_{jn}$

MC for Simple v. Simple

Observe that Type I error and power can be written as **expectations**:

$$P(T \in \mathcal{R} \mid H_0) = E(I(T \in \mathcal{R}) \mid H_0), \quad P(T \in \mathcal{R} \mid H_1) = E(I(T \in \mathcal{R}) \mid H_1)$$

Conditioning on H_0 being true means that $X_i \sim F_0$ (likewise for H_1):

- Generate k **samples of size** n **from** F_0: $X_{j1}, X_{j2}, \ldots, X_{jn}$
- For $j = 1, \ldots k$, **compute** $T_j = T(X_{j1}, \ldots, X_{jn})$ (null distribution)

MC for Simple v. Simple

Observe that Type I error and power can be written as **expectations**:

$$P(T \in \mathcal{R} \mid H_0) = E(I(T \in \mathcal{R}) \mid H_0), \quad P(T \in \mathcal{R} \mid H_1) = E(I(T \in \mathcal{R}) \mid H_1)$$

Conditioning on H_0 being true means that $X_i \sim F_0$ (likewise for H_1):

- Generate k **samples of size** n **from** F_0: $X_{j1}, X_{j2}, \ldots, X_{jn}$
- For $j = 1, \ldots k$, **compute** $T_j = T(X_{j1}, \ldots, X_{jn})$ (null distribution)
- For rejection region $\mathcal{R}$, **estimate** $P(T \in \mathcal{R})$ with the mean of $I(T_j \in \mathcal{R})$

MC for Simple v. Simple

Observe that Type I error and power can be written as **expectations**:

$$P(T \in \mathcal{R} \mid H_0) = E(I(T \in \mathcal{R}) \mid H_0), \quad P(T \in \mathcal{R} \mid H_1) = E(I(T \in \mathcal{R}) \mid H_1)$$

Conditioning on H_0 being true means that $X_i \sim F_0$ (likewise for H_1):

- Generate k **samples of size** n **from** F_0: $X_{j1}, X_{j2}, \ldots, X_{jn}$
- For $j = 1, \ldots k$, **compute** $T_j = T(X_{j1}, \ldots, X_{jn})$ (null distribution)
- For rejection region $\mathcal{R}$, **estimate** $P(T \in \mathcal{R})$ with the mean of $I(T_j \in \mathcal{R})$
- Repeat to get **power** by generating samples from F_1.

Example: Binomial Distribution

Suppose we had some IID data that came from a binomial distribution (I'm hiding the θ parameter for now) of 5 trials.

```
> (xs <- rbinom(10, size = 5, p = hidden_p))

 [1] 0 1 0 3 2 2 0 2 0 1
```

Let's test the **null hypothesis** that $\theta = 0.5$ against the **alternative** that $\theta = 0.6$.

Picking A Test Statistic T

Recall that if $X \sim \text{Binomial}(5, \theta)$, $\mathsf{E}(X) = 5\theta$.

Picking A Test Statistic T

Recall that if $X \sim \text{Binomial}(5, \theta)$, $\mathrm{E}(X) = 5\theta$.

Then the **sample mean divided by 5** is a good estimator of θ.

Recall that if $X \sim \text{Binomial}(5, \theta)$, $\mathrm{E}(X) = 5\theta$.

Then the **sample mean divided by 5** is a good estimator of θ.

We know that in **large samples**, the sample mean is **approximately normal**, but we only have 10 observations. Is that large enough?

Recall that if $X \sim \text{Binomial}(5, \theta)$, $\mathrm{E}(X) = 5\theta$.

Then the **sample mean divided by 5** is a good estimator of θ.

We know that in **large samples**, the sample mean is **approximately normal**, but we only have 10 observations. Is that large enough?

(We also can figure out that the sample mean is **exactly a scaled binomial** $\bar{X} \sim \frac{1}{10}\textbf{Binomial}(10 \times 5, \theta)$, but let's pretend we don't know that.)

Distribution of T when H_0 is true

Key aspects of the data according to H_0:

- 10 observations, IID
- $X_i \sim \text{Binomial}(5, p = \theta)$
- Test statistic value:

$$T = \frac{1}{5n} \sum_{i=1}^{10} X_i$$

Distribution of T when H_0 is true

Key aspects of the data according to H_0:

- 10 observations, IID
- $X_i \sim \text{Binomial}(5, p = \theta)$
- Test statistic value:

$$T = \frac{1}{5n} \sum_{i=1}^{10} X_i$$

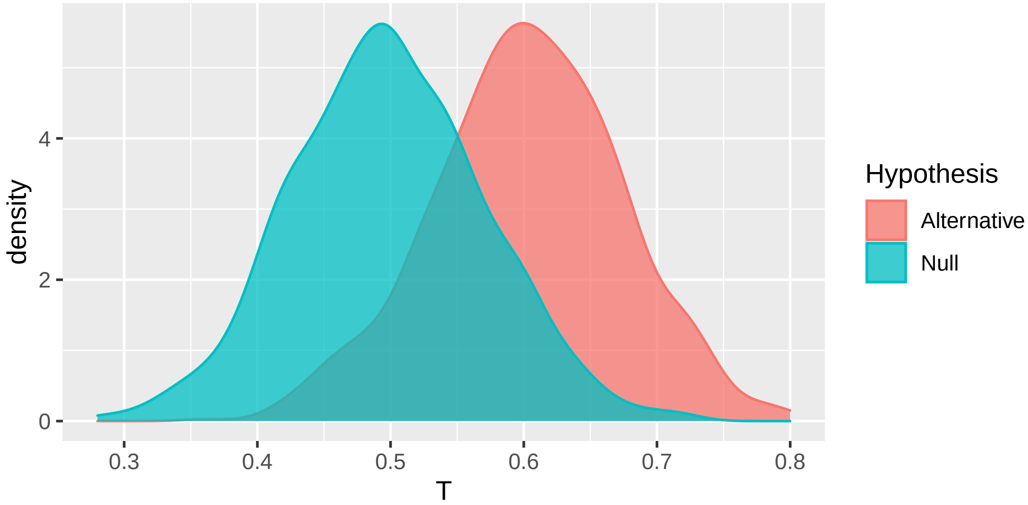
Generate random T values:

```
> k <- 1000
> T_statistic <- function(sample) { mean(sample) / 5 }
> ts_0 <- replicate(k, rbinom(10, size = 5, p = 0.5) %>% T_statistic)
```

Distribution of T when H_1 is true

The only thing that differs is the value of θ:

```
> ts_1 <- replicate(k, rbinom(10, size = 5, p = 0.6) %>% T_statistic)
```


density
4
2
0
0.3
0.4
0.5
0.6
0.7
0.8
T
Hypothesis
Alternative
Null

Picking a rejection region

We want to perform a test with $\alpha = 0.05$. Of any region that has probability of 0.05 under the null, we saw that the **right tail** is probably where most power is:

```
> (rr <- c(quantile(ts_0, 0.95), 1))
 95%
0.62 1.00
```

Picking a rejection region

We want to perform a test with $\alpha = 0.05$. Of any region that has probability of 0.05 under the null, we saw that the **right tail** is probably where most power is:

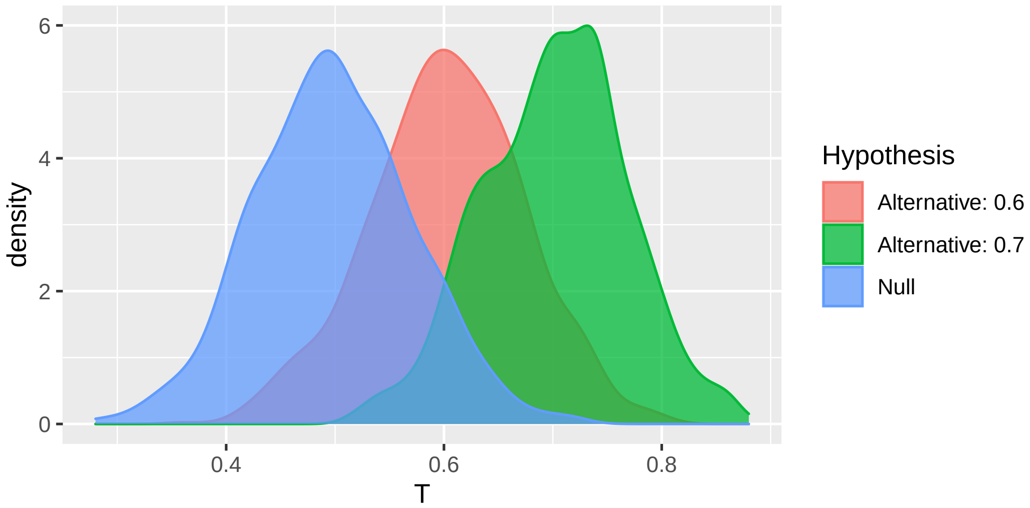
```
> (rr <- c(quantile(ts_0, 0.95), 1))

 95%
0.62 1.00

> T_statistic(xs) # the observed sample

[1] 0.22
```

Since it is less than r_1, **accept (fail to reject)** H_0 **at the** $\alpha = 0.05$ **level**.



density
6
4
2
0
0.4
0.6
0.8
T
Hypothesis
Alternative: 0.6
Alternative: 0.7
Null

Composite Hypotheses

In the previous slide we saw that it seemed for any alternative with $\theta > 0.5$, the best rejection region was in the **right tail of the null distribution**.

In the previous slide we saw that it seemed for any alternative with $\theta > 0.5$, the best rejection region was in the **right tail of the null distribution**.

Or in other words, we would pick the same rejection region for any $H_1 : \theta \in (0.5, 1)$.

Composite Hypotheses

In the previous slide we saw that it seemed for any alternative with $\theta > 0.5$, the best rejection region was in the **right tail of the null distribution**.

Or in other words, we would pick the same rejection region for any $H_1 : \theta \in (0.5, 1)$.

We call such hypotheses **composite** because they contain many simple hypotheses. For example,

$$H_0 : \theta = \theta_0, \quad H_1 : \theta > \theta_0$$

Composite Null Hypotheses

Null hypotheses can also be composite. We just need to be careful about picking a rejection region that has **proper size for any member hypothesis**.

$$\left[\sup_{\theta_0 \in \Theta_0} P(T \in \mathcal{R} \mid \theta = \theta_0)\right] \leq \alpha$$

(sup is "supremum", which we can think of like "maximum", Θ_0 is the set of all null hypotheses).

Composite Null Hypotheses

Null hypotheses can also be composite. We just need to be careful about picking a rejection region that has **proper size for any member hypothesis**.

$$\left[\sup_{\theta_0 \in \Theta_0} P(T \in \mathcal{R} \mid \theta = \theta_0)\right] \leq \alpha$$

(sup is "supremum", which we can think of like "maximum", Θ_0 is the set of all null hypotheses).

In other words, no matter which exact $\theta_0 \in \Theta_0$ holds, we have **size less than level**.

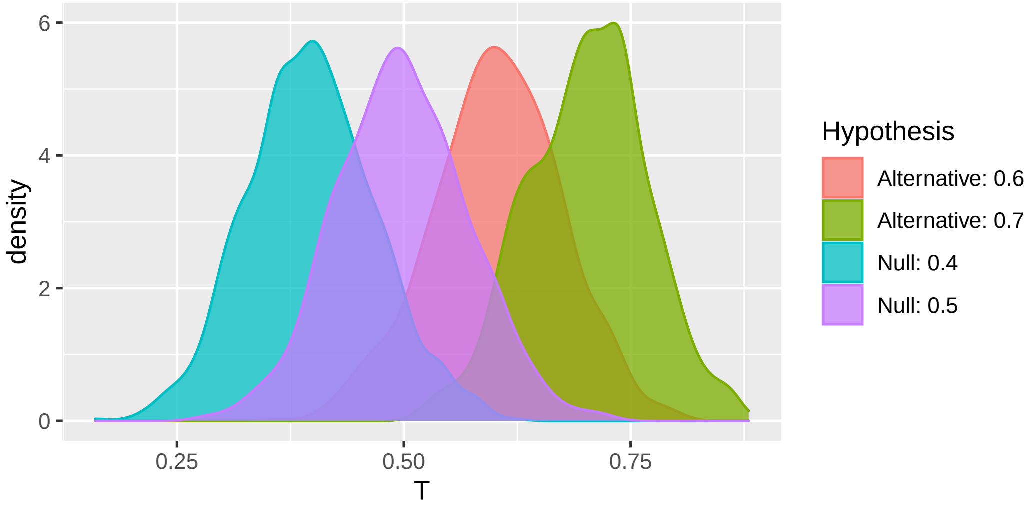
Composite Null Hypotheses

Null hypotheses can also be composite. We just need to be careful about picking a rejection region that has **proper size for any member hypothesis**.

$$\left[\sup_{\theta_0 \in \Theta_0} P(T \in \mathcal{R} \mid \theta = \theta_0)\right] \leq \alpha$$

(sup is "supremum", which we can think of like "maximum", Θ_0 is the set of all null hypotheses).

In other words, no matter which exact $\theta_0 \in \Theta_0$ holds, we have **size less than level**.


6
4
2
0
density
0.25
0.50
0.75
T
Hypothesis
Alternative: 0.6
Alternative: 0.7
Null: 0.4
Null: 0.5

One Tailed Test for Binomial

For $H_0 : \theta = 0.5 vs. H_1 : \theta = 0.6$, we found a rejection region with good power:

```
> rr

 95%
0.62 1.00
```

One Tailed Test for Binomial

For $H_0 : \theta = 0.5 vs. H_1 : \theta = 0.6$, we found a rejection region with good power:

```
> rr

 95%
0.62 1.00
```

We saw this region had good power for any $H_1 : \theta > 0.5$ and also had increasingly smaller size for $H_0 : \theta \leq 0.5$.

One Tailed Test for Binomial

For $H_0 : \theta = 0.5 vs. H_1 : \theta = 0.6$, we found a rejection region with good power:

```
> rr

 95%
0.62 1.00
```

We saw this region had good power for any $H_1 : \theta > 0.5$ and also had increasingly smaller size for $H_0 : \theta \leq 0.5$.

In other words, we have a test of the composite hypothesis:

$$H_0 : \theta \leq 0.5 \quad \text{v.s.} \quad H_1 : \theta > 0$$

From one tailed to two tailed tests

For a parameter $\theta \in (-\infty, \infty)$, consider a hypothesis test of the form:

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta \in (-\infty, \theta_0) \cup (\theta_0, \infty)$$

(we often write $\theta \neq \theta_0$ as short hand for H_1)

For a parameter $\theta \in (-\infty, \infty)$, consider a hypothesis test of the form:

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta \in (-\infty, \theta_0) \cup (\theta_0, \infty)$$

(we often write $\theta \neq \theta_0$ as short hand for H_1)

For **many test statistics**, we can find good rejection regions of the form:

$$(-\infty, T_0(\alpha/2)) \cup (T_0(1 - \alpha/2), \infty)$$

which we can think of **the union of rejection regions** of two one tailed tests.

Two tailed test for binomial example

Let's test:

$$H_0 : \theta = 0.5 \quad \text{v.s.} \quad H_1 : \theta \neq 0.5$$

Two tailed test for binomial example

Let's test:

$$H_0 : \theta = 0.5 \quad \text{v.s.} \quad H_1 : \theta \neq 0.5$$

We saw when $H_1 : \theta > 0.5$, our best rejection region was in the upper tail. We will suppose (correctly) that the best rejection region for $H_1 : \theta > 0.5$ is the lower tail.

Two tailed test for binomial example

Let's test:

$$H_0 : \theta = 0.5 \quad \text{v.s.} \quad H_1 : \theta \neq 0.5$$

We saw when $H_1 : \theta > 0.5$, our best rejection region was in the upper tail. We will suppose (correctly) that the best rejection region for $H_1 : \theta > 0.5$ is the lower tail.

Again testing at $\alpha = 0.05$

```
> rr_upper <- quantile(ts_0, 0.975) # cut the alpha level in half
> rr_lower <- quantile(ts_0, 0.025)
> T_statistic(xs) < rr_lower || T_statistic(xs) > rr_upper # do we reject?
```

```
[1] TRUE
```

The "achieved α level": p-value

For the one sided test $H_1 : \theta > 0.5$, we picked our rejection region as $[\,T_0(1-\alpha), 1)$.

The "achieved α level": p-value

For the one sided test $H_1 : \theta > 0.5$, we picked our rejection region as $[T_0(1-\alpha), 1)$.

What α level would I have to pick to get a rejection region of $[t, 1)$, where t is the observed value of the test statistic?

The "achieved α level": p-value

For the one sided test $H_1 : \theta > 0.5$, we picked our rejection region as $[T_0(1-\alpha), 1)$.

What α level would I have to pick to get a rejection region of $[t, 1)$, where t is the observed value of the test statistic?

$$\alpha \geq P(T \in [t, 1) \mid H_0)$$

The "achieved α level": p-value

For the one sided test $H_1 : \theta > 0.5$, we picked our rejection region as $[T_0(1-\alpha), 1)$.

What α level would I have to pick to get a rejection region of $[t, 1)$, where t is the observed value of the test statistic?

$$\alpha \geq P(T \in [t, 1) \mid H_0) = P(T \geq t \mid H_0) = p$$

The "achieved α level": p-value

For the one sided test $H_1 : \theta > 0.5$, we picked our rejection region as $[T_0(1-\alpha), 1)$.

What α level would I have to pick to get a rejection region of $[t, 1)$, where t is the observed value of the test statistic?

$$\alpha \geq P(T \in [t, 1) \mid H_0) = P(T \geq t \mid H_0) = p \approx \frac{\sum_{j=1}^{k} I(T_j \geq t)}{k}$$

The “achieved α level”: *p*-value

For the one sided test $H_1 : \theta > 0.5$, we picked our rejection region as $[T_0(1-\alpha), 1)$.

What α level would I have to pick to get a rejection region of $[t, 1)$, where t is the observed value of the test statistic?

$$\alpha \geq P(T \in [t, 1) \mid H_0) = P(T \geq t \mid H_0) = p \approx \frac{\sum_{j=1}^{k} I(T_j \geq t)}{k}$$

Sometimes phrased as, “what proportion of T is more extreme than t?”

Duality between p-values and rejecting hypotheses

The p-value tells us the **smallest α level** that would place t in the rejection region.

Duality between p-values and rejecting hypotheses

The p-value tells us the **smallest α level** that would place t in the rejection region.

Implication: For any $\alpha > p$, the rejection region would also contain t (reject). For any $\alpha < p$, t would not be in the rejection region (accept).

Duality between p-values and rejecting hypotheses

The p-value tells us the **smallest α level** that would place t in the rejection region.

Implication: For any $\alpha > p$, the rejection region would also contain t (reject). For any $\alpha < p$, t would not be in the rejection region (accept).

Suppose $\alpha = 0.05$. If $p < 0.05$, then **I would reject that null hypothesis at the 0.05-level**.

The $p-value$ for two-sided tests

We saw for **two sided tests** the rejection region is composed of the union of two one tailed tests.

The $p-value$ for two-sided tests

We saw for **two sided tests** the rejection region is composed of the union of two one tailed tests.

We usually define the two tailed p-value in a similar way:

$$p^{+} = P(T \geq t \mid H_0), p^{-} = P(T \leq t \mid H_0), p\text{-value} = 2\min(p^{+}, p^{-})$$

The $p-value$ for two-sided tests

We saw for **two sided tests** the rejection region is composed of the union of two one tailed tests.

We usually define the two tailed p-value in a similar way:

$$p^+ = P(T \geq t \mid H_0), p^- = P(T \leq t \mid H_0), p\text{-value} = 2\min(p^+, p^-)$$

We can think about this as looking in both tails, but then penalizing ourselves using our data twice.

Computing the p-value

Observed value of the test statistic:

```
> (observed_t <- T_statistic(xs))
[1] 0.22
```

Computing the p-value

Observed value of the test statistic:

```
> (observed_t <- T_statistic(xs))
```

```
[1] 0.22
```

Compute the two one-sided p-values and combine:

```
> p_less    <- mean(ts_0 <= observed_t)
> p_greater <- mean(ts_0 >= observed_t)
> (pvalue <- 2 * min(p_less, p_greater))
```

```
[1] 0
```

Testing $H_0 : \theta = 0.25$

Let's repeat for a different null hypothesis $H_0 : \theta = 0.25$ versus $H_1 : \theta \neq 0.25$.

```
> ts_0.25 <- replicate(10000, T_statistic(rbinom(10, size = 5, p = 0.25)))
> (p_0.25 <- 2 * min(mean(ts_0.25 <= observed_t),
+                     mean(ts_0.25 >= observed_t)))
```

```
[1] 0.7614
```

Testing $H_0 : \theta = 0.25$

Let's repeat for a different null hypothesis $H_0 : \theta = 0.25$ versus $H_1 : \theta \neq 0.25$.

```
> ts_0.25 <- replicate(10000, T_statistic(rbinom(10, size = 5, p = 0.25)))
> (p_0.25 <- 2 * min(mean(ts_0.25 <= observed_t),
+                    mean(ts_0.25 >= observed_t)))
```

```
[1] 0.7614
```

Big reveal: The true θ was 0.25!

Monte Carlo Hypothesis Testing Summary

- Define the **null hypothesis** and **alternative hypothesis** that define distributions for the sample $X_1, \ldots, X_n$.
- Select a **test statistic** $T(X_1, \ldots, X_n)$
- Use the null and alternative hypotheses to draw from $X_1, \ldots, X_n$ and compute T (**null distribution** and **alternative distribution**).
- Find a **rejection region** (subset of support of null distribution) with **size less than specified level**: $P(T \in \mathcal{R} \mid H_0) \leq \alpha$ and good **power**: $P(T \in \mathcal{R} \mid H_1)$
- Compare observed statistic to rejection region or compute a p-value

Extended Example: Benford's Law

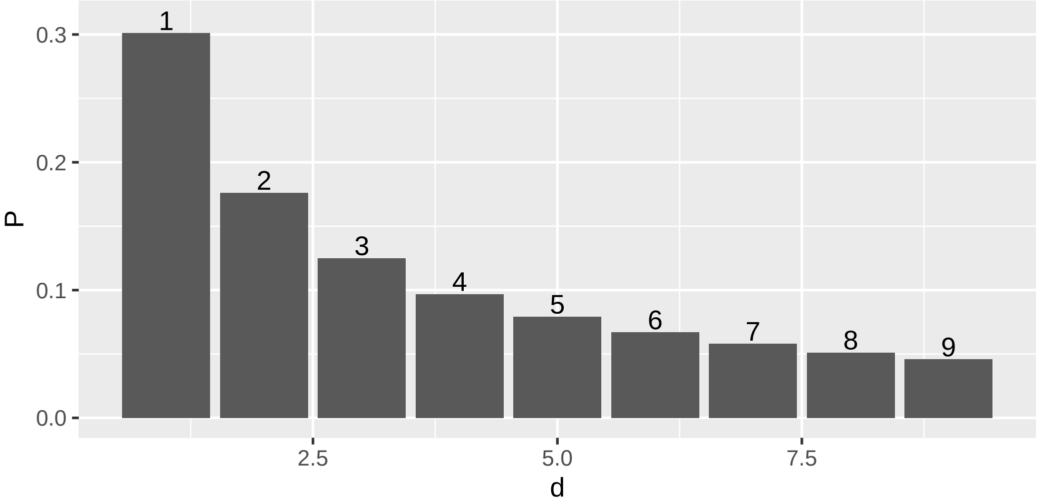
Example: Benford's Law

Benford's Law holds that the distribution of **leading digits** in a collection of numbers spanning several orders of magnitudes will follow the following distribution:

$$\Pr(D = d) = \log_{10}\left(\frac{d+1}{d}\right), \quad d = 1, \dots, 9$$

```
> dbenford <- function(x) {
+     ifelse(x >= 1 & x <= 9, log((x + 1)/ x, base = 10), 0)
+ }
```

$\Pr(D = d)$ under Benford's Law



Using random *D*s

Tam Cho and Gaines (2007) investigated **political contributions between political committees** as reported by the FEC. Here are the digit frequencies for 8,396 contributions in 2004 (Table 1):

```
> pol_digits <- c(23.3, 21.1, 8.5, 11.7, 9.5, 4.2, 3.7, 4.0, 14.1) / 100
```

Tam Cho and Gaines (2007) investigated **political contributions between political committees** as reported by the FEC. Here are the digit frequencies for 8,396 contributions in 2004 (Table 1):

```
> pol_digits <- c(23.3, 21.1, 8.5, 11.7, 9.5, 4.2, 3.7, 4.0, 14.1) / 100
```

A typical way to analyze these data would be to use a χ^2 test comparing the **expected** to the **observed counts**. Alternatively, Tam Cho and Gaines suggest the statistic:

```
> distance <- function(v)  { sqrt(sum((v - dbenford(1:9))^2)) }
```

Hypothesis Test

We will test the null hypothesis that Benford's Law holds for political contributions versus the alternative that it does not hold (goodness-of-fit test).

Hypothesis Test

We will test the null hypothesis that Benford's Law holds for political contributions versus the alternative that it does not hold (goodness-of-fit test).

If the null hypothesis is true, then the observed distance statistic should be close to zero:

```
> (observed_dist <- distance(pol_digits))

[1] 0.1355
```

Hypothesis Test

We will test the null hypothesis that Benford's Law holds for political contributions versus the alternative that it does not hold (goodness-of-fit test).

If the null hypothesis is true, then the observed distance statistic should be close to zero:

```
> (observed_dist <- distance(pol_digits))
```

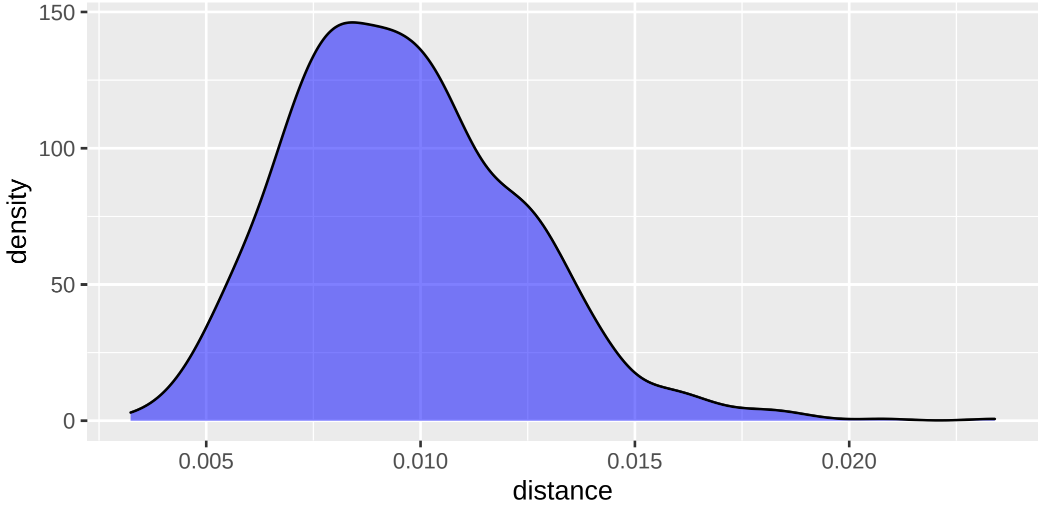
```
[1] 0.1355
```

What is the probability of observing a distance of *at least* 0.1355 if the null hypothesis (Benford's Law) holds?

Distribution of the Test Statistic

```
> rbenford <- function(n) {
+   sample(1:9, size = n, prob = dbenford(1:9), replace = TRUE)
+ }
> n <- 8396
> compute_test_statistic <- function(ds) {
+     probs <- hist(ds, breaks = 0:9, plot = FALSE)$density
+     distance(probs)
+ }
> null_distances <- replicate(1000,
+                               compute_test_statistic(rbenford(n)))
```

Null Distribution



Understanding power of distance test statistic

In order to **pick a rejection region** and **compute the power** of a test statistic, we need to carefully define an alternative hypothesis.

Understanding power of distance test statistic

In order to **pick a rejection region** and **compute the power** of a test statistic, we need to carefully define an alternative hypothesis.

One natural choice would be that **digits are uniformly distributed**: $P(D = d) = 1/9$.

Understanding power of distance test statistic

In order to **pick a rejection region** and **compute the power** of a test statistic, we need to carefully define an alternative hypothesis.

One natural choice would be that **digits are uniformly distributed**: $P(D = d) = 1/9$.

Can we add a parameter θ that controls **how close** to either Bedford or uniform a distribution on digits is?

Parameterizing Alternative

Notice that as $\theta \to \infty$,

$$\frac{d+1+\theta}{d+\theta} \to 1$$

which suggests a model like:

$$P(D=d) = \log_{10}\left(a\frac{d+1+\theta}{d+\theta}\right)$$

Parameterizing Alternative

Notice that as $\theta \to \infty$,

$$\frac{d+1+\theta}{d+\theta} \to 1$$

which suggests a model like:

$$P(D = d) = \log_{10}\left(a\frac{d+1+\theta}{d+\theta}\right)$$

We need to find the normalizing constant a.

Finding a

To get a,

$$\sum_{d=1}^{9} \log_{10}\left(a\frac{d+\theta+1}{d+\theta}\right) = 1 \Rightarrow a^9 \prod_{d=1}^{9} \frac{d+\theta+1}{d+\theta} = 10$$

Finding a

To get a,

$$\sum_{d=1}^{9} \log_{10}\left(a\frac{d+\theta+1}{d+\theta}\right) = 1 \Rightarrow a^9 \prod_{d=1}^{9} \frac{d+\theta+1}{d+\theta} = 10$$

Investigating farther, we see

$$a^9 \frac{(10+\theta)(9+\theta)\cdots(2+\theta)}{(9+\theta)(8+\theta)\cdots(1+\theta)} = a^9 \frac{10+\theta}{1+\theta} = 10$$

so

$$a = \left[\frac{10(1+\theta)}{10+\theta}\right]^{1/9}$$

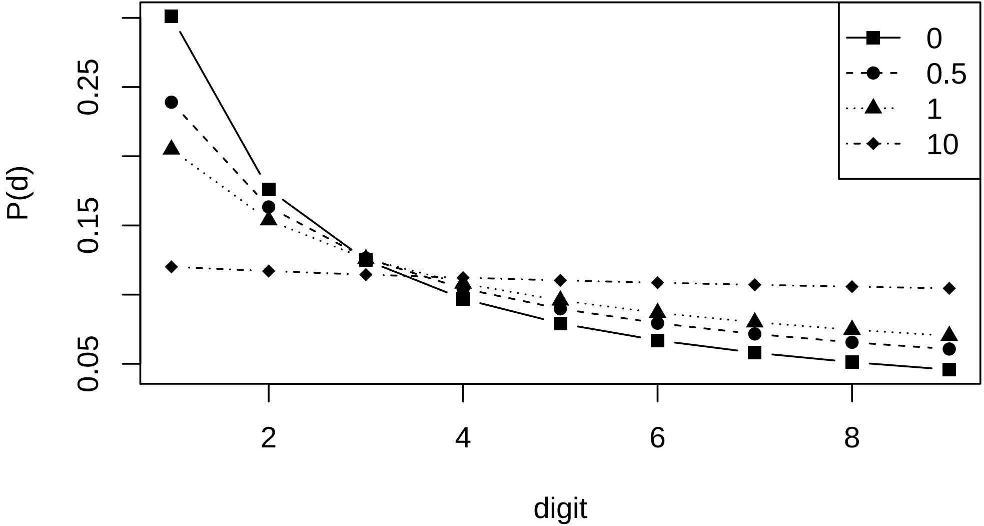
Putting it together

$$P(D = d) = \log_{10}\left(\left[\frac{10(1+\theta)}{10+\theta}\right]^{\frac{1}{9}} \frac{d+\theta+1}{d+\theta}\right), \theta > 0$$

Putting it together

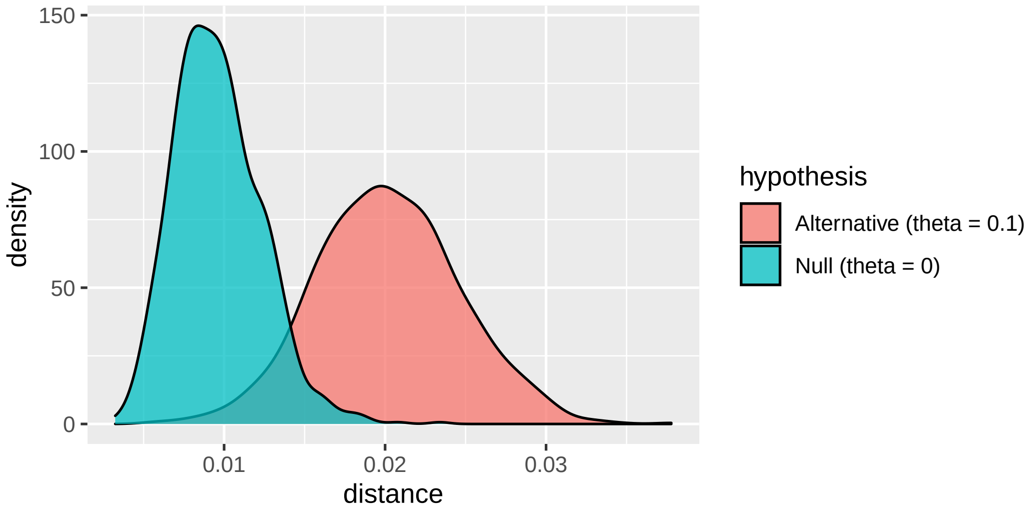
$$P(D = d) = \log_{10}\left(\left[\frac{10(1+\theta)}{10+\theta}\right]^{\frac{1}{9}} \frac{d+\theta+1}{d+\theta}\right), \theta > 0$$

```
> alt_dist <- function(theta) {
+   a <- (10 * (1 + theta) / (10 + theta))^(1/9)
+   log10(a * (1:9 + theta + 1) / (1:9 + theta))
+ }
```


0
0.5
1
10
P(d)
0.05
0.15
0.25
2
4
6
8
digit

Alternative distribution $\theta = 0.1$

```
> alt_0.1 <- replicate(1000, {
+     a_sample <- sample(1:9, size = n, replace = TRUE,
+                        prob = alt_dist(0.1))
+     compute_test_statistic(a_sample)
+ })
>
```


150
100
50
0
density
0.01
0.02
0.03
distance
hypothesis
Alternative (theta = 0.1)
Null (theta = 0)

p-value for the hypothesis test

```
> (p_value <- mean(null_distances >= observed_dist)) # P(T > t)
[1] 0
```

The observed test statistic was larger than any sample we generated (so the p-value was zero) and is 47 standard deviations from the mean of the null distribution.

p-value for the hypothesis test

```
> (p_value <- mean(null_distances >= observed_dist)) # P(T > t)
[1] 0
```

The observed test statistic was larger than any sample we generated (so the *p*-value was zero) and is 47 standard deviations from the mean of the null distribution.

With *extremely high confidence*, we can reject the null hypothesis that these data were a sample from a population that follows Benford's Law.

Power at $\alpha = 0.001$ and $\theta = 0.1$

First, we need to find the 99% quantile under the null:

```
> (rejection_cutoff <- quantile(null_distances, 0.999))

 99.9%
0.0208

> mean(alt_0.1 >= rejection_cutoff)

[1] 0.438
```

Power Curves

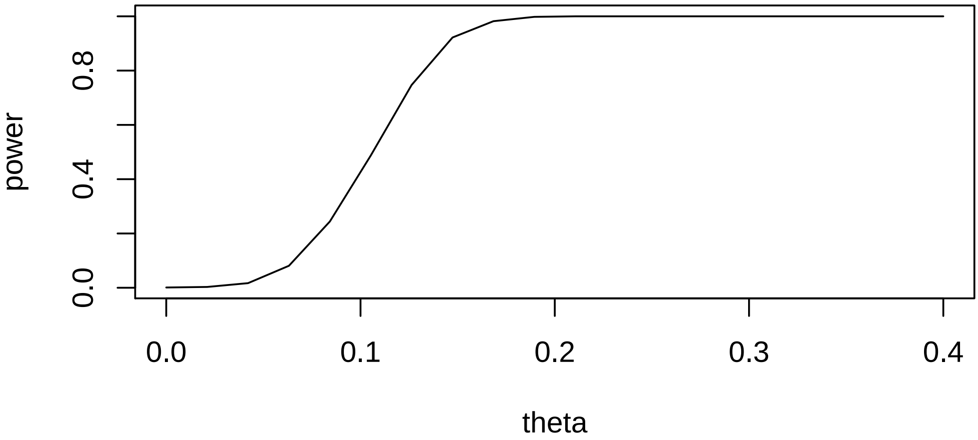
We saw power at one particular point $\theta = 0.1$. What about **other values of θ**?

Power Curves

We saw power at one particular point $\theta = 0.1$. What about **other values of θ**?

We can compute power for many values of θ (holding our α level fixed) to see how it changes.

```
> thetas <- seq(0, 0.4, length.out = 20)
> power_curve <- map_dbl(thetas, function(theta) {
+     alt <- replicate(1000, {
+         a_sample <- sample(1:9, size = n,
+                            replace = TRUE, prob = alt_dist(theta))
+         compute_test_statistic(a_sample)
+     })
+     mean(alt >= rejection_cutoff)
+ })
```


power
0.0
0.4
0.8
0.0
0.1
0.2
0.3
0.4
theta

Future Directions and More on Benford's Law

In this analysis, we considered one possible alternative and one test statistic:

- Why would contributions not follow Benford's law? How could this be turned into an alterantive distribution?
- What other test statistics are possible? What would the power of those statistics be?

Future Directions and More on Benford's Law

In this analysis, we considered one possible alternative and one test statistic:

- Why would contributions not follow Benford's law? How could this be turned into an alterantive distribution?
- What other test statistics are possible? What would the power of those statistics be?

If you are interested in learning more about Benford's Law,

- A Simple Explanation of Benford's Law by R. M. Fewster.
- Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance by Wendy Tam Cho and Brian Gaines