Hypothesis Testing - First Examples

M&Ms Example

We take a sample of 20 M&M's and want to use this sample to learn about the proportion of M&M's that are blue. Let the random variable X denote the number of M&M's in our sample that are blue. We observe $x = 7$.

Statistical Model: $X \sim \text{Binomial}(20, \theta)$

Interpretation of θ: proportion of M&M's that are blue in "population" of all M&M's.

Questions we might ask:

1. What is our best estimate of the proportion of M&M's that are blue based on the observed data x?
   - $\hat{\theta}^{MLE} = \frac{x}{n} = \frac{7}{20}$
   - Method of moments
   - posterior mean, median, or mode.
2. What is a range of plausible values for the proportion of M&M's that are blue based on the observed data x?

   Interval estimate: confidence interval, credible interval

3. As of 2008, the proportion of M&M's that were blue was 0.2.

- Has it changed since then?
- I heard a rumor that it is now 0.25. Is it true?

$\Big\}$ hypothesis tests

Quantify strength of evidence against a specified parameter value (0.2 in this case)

Part 1: Simple Vs. Composite Hypotheses

Definitions and set up

Parameter space:

- Definition and notation: The parameter space Ω is the set of possible values of a parameter.

- In our example: Model $X \sim \text{Binomial}(20, \theta)$

  $\Omega = [0, 1]$

- Set up for hypotheses: Specify subsets of Ω for each hypothesis

  Ω_0 is a subset of Ω specifying values of θ consistent with H_0. Similarly for Ω_A.

  A hypothesis is "simple" if Ω_0 (or Ω_A) contains only a single number.

  It is composite if Ω_0 (or Ω_A) contains more than one number

- Example 1: Has the proportion of M&M's that are blue changed since 2008, when it was 0.2?

  Null hypothesis $H_0 : \theta \in \Omega_0$ where $\underbrace{\Omega_0 = \{0.2\}}_{\text{simple hypothesis}}$

  $(\theta = 0.2)$

  Alternative hypothesis $H_A : \theta \in \Omega_A$ where $\underbrace{\Omega_A = \Omega \setminus \{0.2\}}$

  $(\theta \neq 0.2)$ composite → numbers in $[0, 1]$ other than 0.2.

- Example 2: Has the proportion of M&M's that are blue changed since 2008, from 0.2 to 0.25?

  $H_0 : \theta \in \{0.2\}$ ← Ω_0

  $H_A : \theta \in \{0.25\}$ ← Ω_A

  both are simple hypotheses.

Example 1: Simple vs Complex Hypotheses

####Summary of previous set up:

- Sample size $n = 20$, observed $x = 7$ blue M&Ms
- Our model is $X \sim \text{Binomial}(20, \theta)$
- Our hypotheses are: $H_0 : \theta = 0.2$ and $H_A : \theta \neq 0.2$

$H_0 : \theta \in \{0.2\}$, $H_A : \theta \in \Omega \setminus \{0.2\}$

Test Statistic:

- General idea: A random variable $W = g(X_1, \dots, X_n)$ that can be used to measure how consistent the data are with the null hypothesis. We would like the distribution of W to be different depending on whether H_0 or H_A is correct.

- Test statistic in our example (and its value based on observed data):

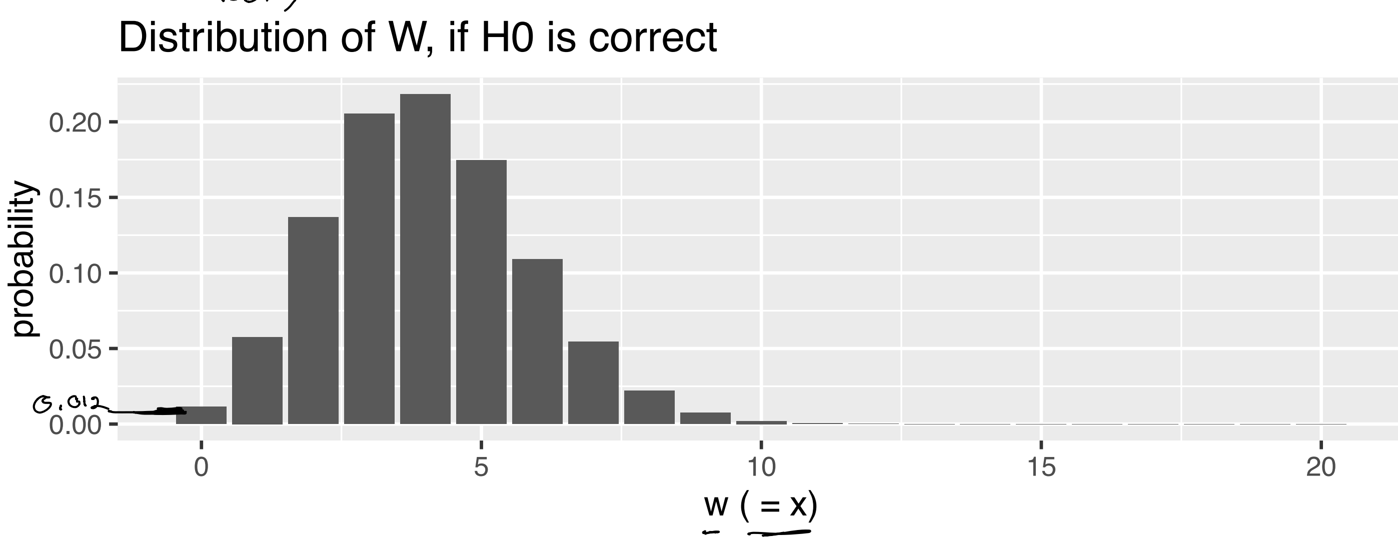
We could use $W = X$

$g(X) = X$, the identity function

Distribution of the test statistic if H_0 is true:

$W \sim \text{Binomial}(20, \theta)$.

If H_0 is correct, $\theta = 0.2$ so $W \sim \text{Binomial}(20, 0.2)$




```
##                 0     1     2     3     4     5     6     7     8     9    10
## probability 0.012 0.058 0.137 0.205 0.218 0.175 0.109 0.055 0.022 0.007 0.002
##                11  12  13  14  15  16  17  18  19  20
## probability     0   0   0   0   0   0   0   0   0   0
```

p-value

- General definition: $P(W$ at least as extreme as $w \mid H_0$ is true$)$

By "at least as extreme", we mean is "at least as inconsistent with H_0".
A small p-value means the observed w is not consistent with H_0, so provides some evidence against H_0.

- In our example: $P(W$ is at least as extreme as $7 \mid \theta = 0.2)$

observed $w=7$ blue M&M's

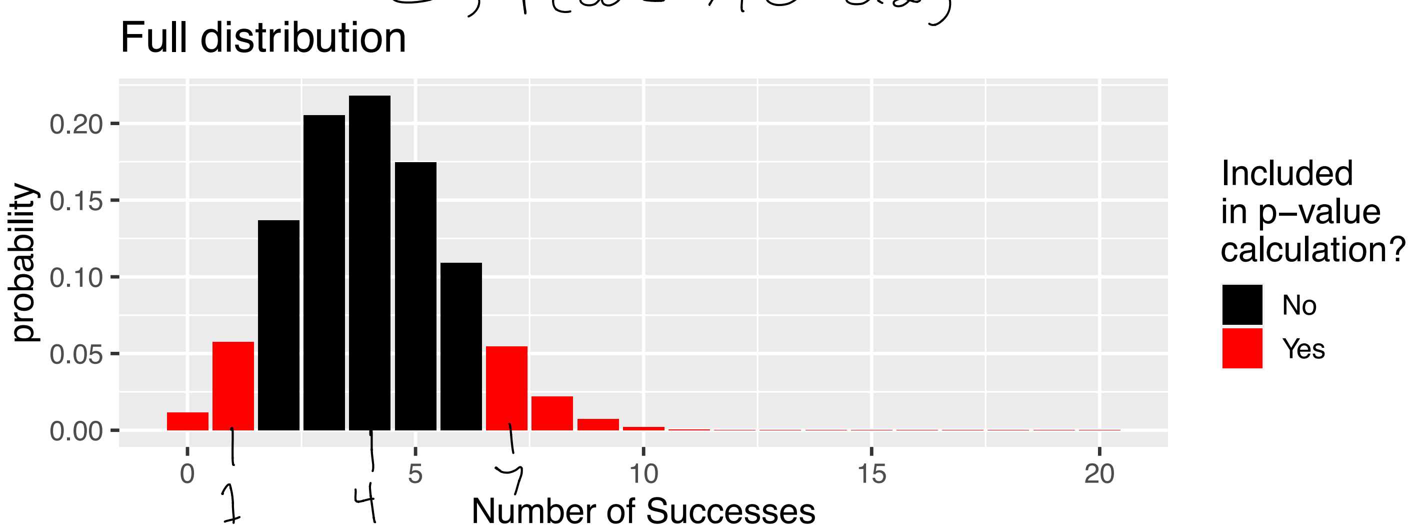
We will interpret "at least as extreme" as meaning at least as far from $E[W]$, if $\theta = 0.2$.
If $\theta = 0.2$, then $E(W) = 20 \cdot 0.2 = 4$.
p-value $= P(W$ is at least as far from 4 as 7 is $\mid \theta = 0.2)$
$= P(W \leq 1$ or $W \geq 7 \mid \theta = 0.2)$

```
pbinom(q = 1, size = 20, prob = 0.2) +
  pbinom(q = 7, size = 20, prob = 0.2, lower.tail = FALSE)
```

$\rightarrow P(W \leq 1 \mid \theta = 0.2)$

$\rightarrow P(W \geq 7 \mid \theta = 0.2)$

```
## [1] 0.101318
```



Sample size $n = 541$, observed $x = 138$ blue M&Ms

- If H_0 is true, then $X \sim \text{Binomial}(541, 0.2)$
- The p-value is $P(X$ at least as extreme as $138)$ given that $X \sim \text{Binomial}(541, 0.2)$
  - $E(X) = 541 * 0.2 = 108.2$
  - $138 - 108.2 = 29.8$
  - $108.2 - 29.8 = 78.4$
  - $P(X$ at least as extreme as $138) = P(X \leq 78$ or $X \geq 138)$

Example 2: Simple Hypotheses

Summary of previous set up:

- Sample size $n = 20$, observed $x = 7$ blue M&Ms
- Our model is $X \sim \text{Binomial}(20, \theta)$
- I heard a rumor that the proportion of M&Ms that are blue was changed to 25. Is it true??
- Our hypotheses are: $H_0 : \theta = 0.2$ and $H_A : \underline{\theta = 0.25}$

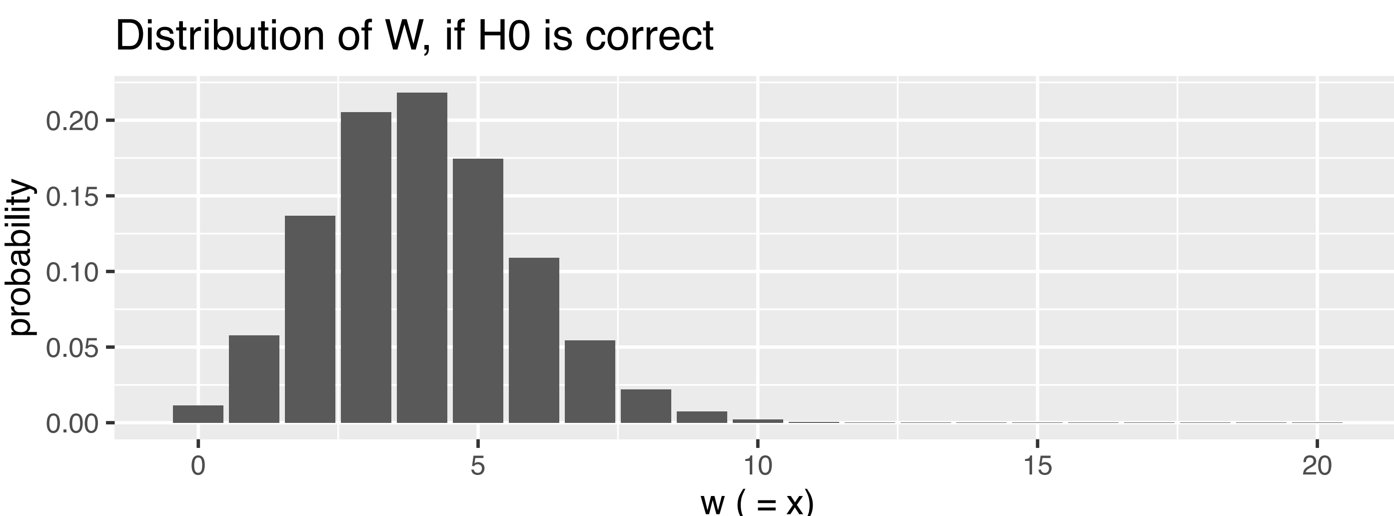
Test statistic

- Test statistic in our example (and its value based on observed data):

$W = X$, observed value is $w = 7$

Distribution of the test statistic if H_0 is true:

$W \sim \text{Binomial}(20, 0.2)$



p-value: $P(W$ is at least as extreme as $7 \mid \theta = 0.2)$

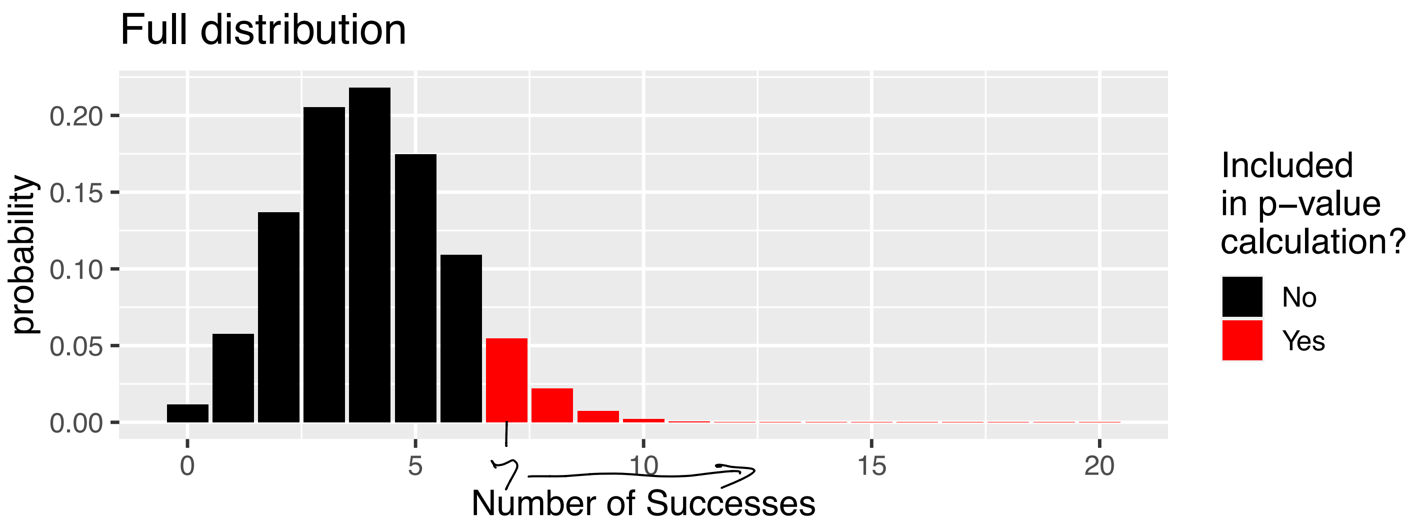
Specifically compare to a larger value of θ, $\theta = 0.25$

Now: "more incompatible with H_0" corresponds to larger W.

$P(W \geq 7 \mid \theta = 0.2)$

```r
pbinom(q = 7, size = 20, prob = 0.2, lower.tail = FALSE)
```

```
## [1] 0.03214266
```




```
##     x probability
## 1   0       0.012
## 2   1       0.058
## 3   2       0.137
## 4   3       0.205
## 5   4       0.218
## 6   5       0.175
## 7   6       0.109
## 8   7       0.055
## 9   8       0.022
## 10  9       0.007
## 11 10       0.002
## 12 11       0.000
## 13 12       0.000
## 14 13       0.000
## 15 14       0.000
## 16 15       0.000
## 17 16       0.000
## 18 17       0.000
## 19 18       0.000
## 20 19       0.000
## 21 20       0.000
```

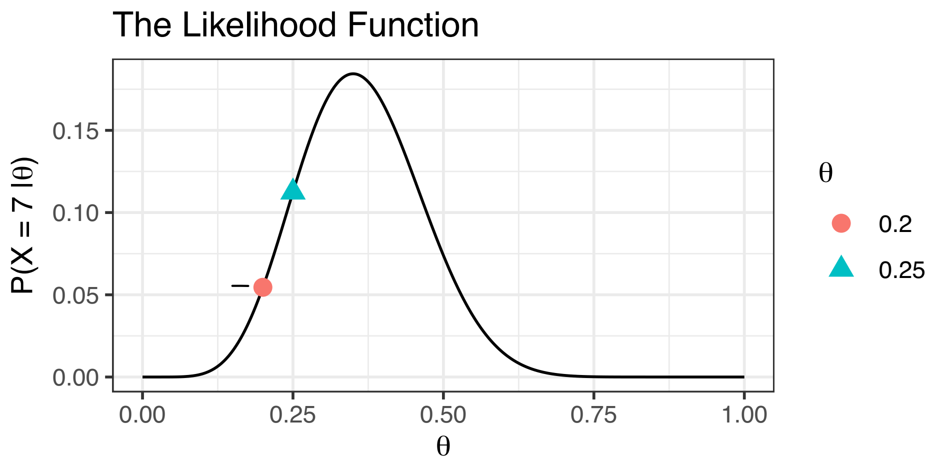
(if $\theta = 0.2$)

p-value is the sum of these numbers

$$\frac{\mathcal{L}(\theta=0.2|X)}{\mathcal{L}(\theta=0.25|X)} = \frac{\text{likelihood if } H_0 \text{ true}}{\text{likelihood if } H_A \text{ is true.}}$$

Another choice for the test statistic

- We can use the statistic $W = g(X) = \frac{f_X(X|\theta=0.2)}{f_X(X|\theta=0.25)}$ (the **likelihood ratio**)




```
w_obs <- dbinom(x = 7, size = 20, prob = 0.2)/dbinom(x = 7, size = 20, prob = 0.25)
w_obs
```

```
## [1] 0.4852922
```

observed likelihood ratio based on sample data $x=7$

Questions:

1. Is a small likelihood ratio or a large likelihood ratio stronger evidence against the null hypothesis?

Small likelihood ratio <=> probability of observed data is small if H_0 true,
" " " " is large if H_A true

<=> evidence against H_0.

2. What should count as "at least as extreme" for the purpose of calculating a p-value based on the likelihood ratio, W?

"at least as extreme" <=> "at least as incompatible with H_0"

<=> at least as small a value of w.

- If H_0 is true, then $X \sim \text{Binomial}(20, 0.2)$
- The p-value is $P(W \leq w)$ given that $X \sim \text{Binomial}(20, 0.2)$

```r
# Manual calculation of the probability distribution of W
x <- seq(from = 0, to = 20)
W_X_distn <- data.frame(
  x = x,
  probability = dbinom(x, size = 20, prob = 0.2),
  w = dbinom(x, size = 20, prob = 0.2) / dbinom(x, size = 20, prob = 0.25)
)

W_X_distn
```

```
##     x  probability          w
## 1   0 1.152922e-02 3.63558642
## 2   1 5.764608e-02 2.72668981
## 3   2 1.369094e-01 2.04501736
## 4   3 2.053641e-01 1.53376302
## 5   4 2.181994e-01 1.15032226
## 6   5 1.745595e-01 0.86274170
## 7   6 1.090997e-01 0.64705627
## 8   7 5.454985e-02 0.48529221
## 9   8 2.216088e-02 0.36396915
## 10  9 7.386959e-03 0.27297687
## 11 10 2.031414e-03 0.20473265
## 12 11 4.616849e-04 0.15354949
## 13 12 8.656592e-05 0.11516212
## 14 13 1.331783e-05 0.08637159
## 15 14 1.664729e-06 0.06477869
## 16 15 1.664729e-07 0.04858402
## 17 16 1.300570e-08 0.03643801
## 18 17 7.650410e-10 0.02732851
## 19 18 3.187671e-11 0.02049638
## 20 19 8.388608e-13 0.01537229
## 21 20 1.048576e-14 0.01152922
```

$$\frac{\mathcal{L}(\theta=0.2 \mid x=0)}{\mathcal{L}(\theta=0.25 \mid x=0)} = \frac{P(X=0 \mid \theta=0.2)}{P(X=0 \mid \theta=0.25)}$$

$$w = \frac{\mathcal{L}(\theta=0.2 \mid x=1)}{\mathcal{L}(\theta=0.25 \mid x=1)} = \frac{P(X=1 \mid \theta=0.2)}{P(X=1 \mid \theta=0.25)}$$

$x=7$, $w = 0.48529221$

$$P(W \leq 0.48529221)$$
$$= P(X \geq 7)$$

```r
# Find the p-value
sum(W_X_distn$probability[W_X_distn$w <= w_obs])
```

```
## [1] 0.08669251
```

Same p-value based on same entries of distribution table.

- Note: in this example, using X or $W = g(X) = \frac{f_X(X|\theta=0.2)}{f_X(X|\theta=0.25)}$ is *equivalent*
  - A larger value of X is less consistent with H_0
  - A smaller value of W is less consistent with H_0

```r
# Find the p-value, if we had used X for our statistic
sum(W_X_distn$probability[W_X_distn$x >= 7])
```

```
## [1] 0.08669251
```

Main point: likelihood ratio test $\left(w = \frac{\mathcal{L}(0.2|x)}{\mathcal{L}(0.25|x)}\right)$ is equivalent to a test based on $w = X$.

To see this, show $W \le w$ if and only if $\underline{X \ge x}$

Verify:

$$\underbrace{W = \frac{\mathcal{L}(0.2 \mid X)}{\mathcal{L}(0.25 \mid X)} \le \boxed{\frac{\mathcal{L}(0.2 \mid x=7)}{\mathcal{L}(0.25 \mid x=7)}}}_{\text{want to show this inequality is equivalent to } X \ge x.}$$

w

$$\Leftrightarrow \frac{\cancel{\binom{n}{X}} 0.2^{X} (1-0.2)^{n-X}}{\cancel{\binom{n}{X}} 0.25^{X} (1-0.25)^{n-X}} \le \frac{\cancel{\binom{n}{x}} 0.2^{x} (1-0.2)^{n-x}}{\cancel{\binom{n}{x}} 0.25^{x} (1-0.25)^{n-x}}$$

$$\Leftrightarrow \left(\frac{0.2}{0.25}\right)^{X} \left(\frac{1-0.2}{1-0.25}\right)^{n-X} \le \left(\frac{0.2}{0.25}\right)^{x} \left(\frac{1-0.2}{1-0.25}\right)^{n-x}$$

$$\Leftrightarrow X \cdot \log\left(\frac{0.2}{0.25}\right) + (\cancel{n} - X)\log\left(\frac{0.8}{0.75}\right) \le x \cdot \log\left(\frac{0.2}{0.25}\right) + (\cancel{n} - x) \cdot \log\left(\frac{0.8}{0.75}\right)$$

$$\Leftrightarrow - X \cdot \log\left(\frac{0.25}{2}\right) + (-X) \cdot \log\left(\frac{0.8}{0.75}\right) \le -x \log\left(\frac{0.25}{0.2}\right) - x \cdot \log\left(\frac{0.8}{0.75}\right)$$

$$\Leftrightarrow -\left\{\log\left(\frac{0.25}{0.2}\right) + \log\left(\frac{0.8}{0.75}\right)\right\} \cdot X \le -\left\{\log\left(\frac{0.25}{0.2}\right) + \log\left(\frac{0.8}{0.75}\right)\right\} \cdot x$$

$$\Leftrightarrow X \ge x$$

How to use Likelihood Ratio Test (LRT)

- Write down likelihood ratio

$$W = \frac{\mathcal{L}(\theta_0 \mid \underline{X})}{\mathcal{L}(\theta_A \mid \underline{X})}, \quad \text{observed } w = \frac{\mathcal{L}(\theta_0 \mid x)}{\mathcal{L}(\theta_A \mid x)}$$

- try to simplify the inequality $W \le w$ to something more intuitive.

- calculate p-value for test.