

2.11. The Maximum of  $n$  Random Variables  
3.4. Hypothesis Testing  
5.4. Long Repeats of the Same Nucleotide

Prof. Tesler

Math 283  
October 21, 2013

# Maximum of two rolls of a die

- Let  $X, Y$  be two rolls of a four sided die and  $U = \max\{X, Y\}$ :

$U$	$X = 1$	2	3	4
$Y = 1$	1	2	3	4
2	2	2	3	4
3	3	3	3	4
4	4	4	4	4

- $$\begin{aligned}P(U = 3) &= F_U(3) - F_U(2) \\&= P(X \leq 3, Y \leq 3) - P(X \leq 2, Y \leq 2) \\&= P(X \leq 3)^2 - P(X \leq 2)^2 \quad (\text{since } X, Y \text{ are i.i.d.}) \\&= F_X(3)^2 - F_X(2)^2\end{aligned}$$
- If it's a fair die then  $F_X(2) = 1/2$ ,  $F_X(3) = 3/4$ , so
$$P(U = 3) = (3/4)^2 - (1/2)^2 = 5/16$$

# Maximum of $n$ i.i.d. random variables: CDF

- Let  $Y_1, \dots, Y_n$  be i.i.d. random variables, each with the same cumulative distribution function  $F_Y(y) = P(Y_i \leq y)$ .
- Let  $Y_{\max} = \max\{Y_1, \dots, Y_n\}$ .
- The cdf of  $Y_{\max}$  is

$$\begin{aligned} F_{Y_{\max}}(y) &= P(Y_{\max} \leq y) \\ &= P(Y_1 \leq y, Y_2 \leq y, \dots, Y_n \leq y) \\ &= P(Y_1 \leq y) P(Y_2 \leq y) \cdots P(Y_n \leq y) \\ &= F_Y(y)^n \end{aligned}$$

# Maximum of $n$ i.i.d. random variables: PDF

## Continuous case

Suppose each  $Y_i$  has density  $f_Y(y)$ . Then  $Y_{\max}$  has density

$$f_{Y_{\max}}(y) = \frac{d}{dy} F_Y(y)^n = n F_Y(y)^{n-1} \frac{d}{dy} F_Y(y) = n F_Y(y)^{n-1} f_Y(y)$$

## Discrete case (integer-valued)

Suppose the random variables  $Y_i$  range over  $\mathbb{Z}$  (integers). For  $y \in \mathbb{Z}$ ,

$$P(Y_{\max} = y) = P(Y_{\max} \leq y) - P(Y_{\max} \leq y-1) = F_Y(y)^n - F_Y(y-1)^n$$

For any non-integer  $y$ ,  $P(Y_{\max} = y) = 0$ .

## Discrete case (in general)

If the random variables  $Y_i$  are discrete and real valued, then for all  $y$ ,

$$P(Y_{\max} = y) = P(Y_{\max} \leq y) - P(Y_{\max} \leq y^-) = F_Y(y)^n - F_Y(y^-)^n$$

# Example: Geometric distribution

(version where  $Y$  counts the number of heads before the first tail)

- $p$  is the probability of heads,  $1 - p$  is the probability of tails.
- Let  $P(Y = y) = p^y(1 - p)$  for  $y = 0, 1, 2, \dots$

- **Cumulative distribution:** For  $y = 0, 1, 2, \dots$ ,

$$\begin{aligned}F_Y(y) &= P(Y \leq y) \\&= p^0(1 - p) + p^1(1 - p) + \dots + p^y(1 - p) \\&= (1 - p) + (p - p^2) + \dots + (p^y - p^{y+1}) \\&= 1 - p^{y+1}\end{aligned}$$

- **Alternate proof:**

- $P(Y \geq y + 1) = p^{y+1}$ :  
there are  $y + 1$  or more heads before the first tails iff the first  $y + 1$  flips are heads.
- $P(Y \leq y) = 1 - p^{y+1}$

# Example: Geometric distribution

## Geometric random variables $Y_1, \dots, Y_n$

- Let  $Y_1, \dots, Y_n$  be i.i.d. geometric random variables, with PDF  $P(Y_i = y) = p^y(1 - p)$  for  $y = 0, 1, 2, \dots$
- **CDF of  $Y_i$ :**  $F_{Y_i}(y) = 1 - p^{y+1}$  for  $y = 0, 1, 2, \dots$

## Distribution of $Y_{\max} = \max\{Y_1, \dots, Y_n\}$

- **CDF of  $Y_{\max}$ :**  $P(Y_{\max} \leq y) = (1 - p^{y+1})^n$  for  $y = 0, 1, 2, \dots$
- **PDF of  $Y_{\max}$ :**

$$\begin{aligned} P(Y_{\max} = y) &= (F_{Y_1}(y))^n - (F_{Y_1}(y-1))^n \\ &= \begin{cases} (1 - p^{y+1})^n - (1 - p^y)^n & \text{if } y = 0, 1, 2, \dots; \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

*Note:* For  $y = 0$ , that's

$$(F_{Y_i}(0))^n - (F_{Y_i}(-1))^n = (1 - p)^n - (1 - 1)^n = (1 - p)^n$$

but  $F_{Y_i}(-1)$  is out of range, so check:

$$P(Y_{\max} = 0) = P(Y_1 = \dots = Y_n = 0) = (1 - p)^n$$

# Related problems

## Minimum

Find the distribution of the minimum of  $n$  i.i.d. random variables.

## Order statistics (Chapter 2.12)

Given random variables  $Y_1, Y_2, \dots, Y_n$ , reorder as  $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ :

- Find the distribution of the 2nd largest (or  $k$ th largest/smallest).
- Find the joint distribution of the 2nd largest and 5th smallest, or any other combination of any number of the  $Y_{(i)}$ 's (including all).

## Applications

- Distribution of the median of repeated indep. measurements.
- Cut up genome by a Poisson process (crossovers; restriction fragments; genome rearrangements), put the fragment lengths into order smallest to largest, and analyze the joint distribution.
- Beta distribution (Ch. 1.10.6): using Gamma distribution notation: distribution of  $D_3/D_8$  (position of 3rd cut as fraction of 8th)?

# Long repeats of the same letter

We consider DNA sequences of length  $N$ , and want to distinguish between two *hypotheses*:

## “Null Hypothesis” $H_0$ :

The DNA sequence is generated by independent rolls of a 4-sided die (A,C,G,T) with probabilities  $p_A, p_C, p_G, p_T$  that add to 1.

## “Alternative Hypothesis” $H_1$ :

Adjacent positions are correlated: there is a tendency for long repeats of the letter A.

- We will develop a quantitative way to determine whether  $H_0$  or  $H_1$  better applies to a sequence.
- We will cover a number of other hypothesis tests in this class.



# Longest run of A's in a sequence

- Split a sequence after every non-A:

T/AAG/AC/AAAG/G/T/C/AG/

- Let  $Y_1, \dots, Y_n$  be the number of A's in each segment, and let  $Y_{\max} = \max\{Y_1, \dots, Y_n\}$ :

$\underbrace{T}_{y_1=0} / \underbrace{AAG}_{y_2=2} / \underbrace{AC}_{y_3=1} / \underbrace{AAAG}_{y_4=3} / \underbrace{G}_{y_5=0} / \underbrace{T}_{y_6=0} / \underbrace{C}_{y_7=0} / \underbrace{AG}_{y_8=1} /$

- $n = 8$  and  $y_{\max} = 3$ .
- We will use  $y_{\max}$  as a *test statistic* to decide if we are more convinced of  $H_0$  or  $H_1$ :
  - Smaller values of  $y_{\max}$  support  $H_0$ .
  - Larger values of  $y_{\max}$  support  $H_1$ .
  - There are clear-cut cases, and a gray zone in-between.

**The null hypothesis,  $H_0$ , is given the benefit of the doubt in ambiguous cases.**

# Hypothesis testing

- 1 State a null hypothesis  $H_0$  and an alternative hypothesis  $H_1$ :
  - $H_0$ : The DNA sequence is generated by independent rolls of a 4-sided die (A,C,G,T) with probabilities  $p_A, p_C, p_G, p_T$ , that add to 1.
  - $H_1$ : Adjacent positions are correlated: there is a tendency for long repeats of the letter A.
- 2 Compute a *test statistic*:  $y_{\max}$ .
- 3 Calculate the *P-value*:  $P = P(Y_{\max} \geq y_{\max})$ .
  - Assuming  $H_0$  is true, what is the probability to observe the test statistic “as extreme or more extreme” as the observed value?
  - “Extreme” means away from  $H_0$  / towards  $H_1$ .
- 4 *Decision*: Does  $H_0$  or  $H_1$  apply?
  - If the  $P$ -value is too small (typically  $\leq 5\%$  or  $\leq 1\%$ ), we *reject the null hypothesis* (Reject  $H_0$ ) / *accept the alternative hypothesis* (Accept  $H_1$ ).
  - Otherwise, we *accept the null hypothesis* (Accept  $H_0$ ) / *reject the alternative hypothesis* (Reject  $H_1$ ).
  - Picky people prefer “Reject  $H_0$ ” vs. “Insufficient evidence to reject  $H_0$ .”

# Computing the $P$ -value

- **$P$ -value:** Assuming  $H_0$  is true, what is the probability to observe a test statistic at least as “extreme” (away from  $H_0$  / towards  $H_1$ ) as the observed test statistic value?
- The  **$P$ -value** in this problem is  $P = P(Y_{\max} \geq y_{\max})$ .
- **Notation:**
  - $p = p_A$  is the probability of  $A$ 's under  $H_0$ ,
  - $N$  = length of the sequence,
  - $n$  = number of runs of  $A$ 's,
  - $y_{\max}$  = number of  $A$ 's in the longest run.
- **Notation peculiarities:**
  - The  $N$  &  $n$  notation does not follow the usual conventions on uppercase/lowercase for random variables vs. specific measurements.
  - The non- $A$ 's have a Binomial( $N, 1 - p$ ) distribution:  
 $N$  positions, each with probability  $1 - p$  not to be an  $A$ .  
Additionally,  $n$  counts the number of the non- $A$ 's, since these terminate the runs of  $A$ 's (including runs of 0  $A$ 's).

# Computing the $P$ -value

- By the Binomial( $N, 1 - p$ ) distribution, approximately  $(1 - p)N$  letters are not  $\mathbb{A}$ , giving an estimate of  $n \approx (1 - p)N$  runs.
- Each run has a geometric distribution (# “heads” before first tails) with parameter  $p$  of “heads” ( $\mathbb{A}$ ):

$$P_{Y_i}(y) = (1 - p)p^y \qquad F_{Y_i}(y) = 1 - p^{y+1}$$

- For an observation  $y = y_{\max} = 0, 1, 2, \dots$ :

$$\begin{aligned} P &= P(Y_{\max} \geq y) = 1 - P(Y_{\max} \leq y - 1) \\ &= 1 - P(Y_1 \leq y - 1)^n = 1 - (F_{Y_1}(y - 1))^n \\ &= 1 - (1 - p^y)^n = 1 - (1 - p^y)^{(1-p)N} \end{aligned}$$

- The table shows  $P$ -values for  $p = p_{\mathbb{A}} = .25$  and sequence length  $N = 100,000$ .

$y_{\max}$	$P$
$\leq 5$	1.
6	0.99999
7	0.98972
8	0.68159
9	0.24881
10	0.06902
11	0.01772
12	0.00446
13	0.00111
14	0.00027
15	0.00006

# Decision

- We will choose a *critical value* or *cutoff*  $y^*$ , and make the decision
  - “Accept  $H_0$ ” (“Accept the null hypothesis”) when  $y_{\max} \leq y^*$ ;  
a.k.a. “Reject  $H_1$ ” (“Reject the alternative hypothesis”) or “Fail to reject  $H_0$ .”
  - “Accept  $H_1$ ” / “Reject  $H_0$ ” when  $y_{\max} > y^*$   
 (“Accept the alternative hypothesis” / “Reject the null hypothesis”)
- How do we choose this critical value?  
There are clear-cut cases, and a gray zone in-between.  
**The null hypothesis  $H_0$  is given the benefit of the doubt in ambiguous cases.**
- If observed data would be very unlikely (even if still possible) under  $H_0$ , we reject  $H_0$  and accept  $H_1$ .
- Choose a *significance level*  $\alpha$  (usually 5% or 1%).
- Determine the critical value so that when  $H_0$  is true, at most a fraction  $\alpha$  of the cases will be misclassified as  $H_1$  (a *Type I error*).
- We will separately consider *Type II errors* (accepting  $H_0$  when  $H_1$  is really true).

# Decision procedure (using a cutoff on the test statistic)

$y_{\max}$	$P$
$\leq 5$	1.
6	0.99999
7	0.98972
8	0.68159
9	0.24881
10	0.06902
11	0.01772
12	0.00446
13	0.00111
14	0.00027
15	0.00006

- Choose a cutoff so that when  $H_0$  is really true, we incorrectly reject  $H_0$  at most a fraction  $\alpha$  of the time.
- $\alpha = .05 = 5\%$ :  
Accept  $H_0$  when  $y_{\max} \leq 10$ ;  
Reject  $H_0$  when  $y_{\max} \geq 11$ .
- When  $H_0$  is true, this incorrectly rejects  $H_0$  (a *Type I error*) a fraction  $0.01772 = 1.772\%$  of the time.
- A continuous test statistic would have a cutoff giving exactly 5%. This one is discrete, so it jumps.
- $\alpha = .01 = 1\%$ :  
Accept  $H_0$  when  $y_{\max} \leq 11$ ;  
Reject  $H_0$  when  $y_{\max} \geq 12$ .  
Type I error rate 0.446%.
- The *Type II error rate* is the fraction of the time that  $H_0$  is accepted when  $H_1$  is really true. We did not formulate  $H_1$  precisely enough to compute it.

# Decision procedure (using $P$ -value instead of $y_{\max}$ )

$y_{\max}$	$P$
$\leq 5$	1.
6	0.99999
7	0.98972
8	0.68159
9	0.24881
10	0.06902
11	0.01772
12	0.00446
13	0.00111
14	0.00027
15	0.00006

- Determine the  $P$ -value of the test statistic.
- Accept  $H_0$  when  $P > \alpha$ ;  
Reject  $H_0$  when  $P \leq \alpha$ .
- This is equivalent to the first decision procedure:  
For  $\alpha = 0.05$ , we have  
 $P > 0.05$  when  $y_{\max} \leq 10$ : Accept  $H_0$   
 $P \leq 0.05$  when  $y_{\max} \geq 11$ : Reject  $H_0$

# Advantages of using $P$ -values over critical values in hypothesis tests

- $P$ -values can be defined for any hypothesis test, so hypothesis tests on other types of data with other types of test statistics can also be formulated as “Reject  $H_0$  if  $P \leq \alpha$ ” instead of in a manner specific to the particular test.
- It's easy to tell if you're near the cutoff when using  $P$ .  
Using the test statistic, you'd have to determine that for each test statistic based on its distribution.  
E.g., is being within  $\pm 100$  close?  $\pm 10$ ?  $\pm 1$ ?  
It all depends on the distribution of the statistic.
- $P$ -values allow you to test against cutoffs for several  $\alpha$ 's simultaneously.



# Example: SARS — Genome sequence

- The complete genome is at

<http://www.ncbi.nlm.nih.gov/nucore/30271926?report=genbank>

- It consists of  $N = 29751$  bases, fully sequenced, no gaps.

Nucleotide	Frequency	Proportion
A	8481	$p_A \approx 0.2851$
C	5940	$p_C \approx 0.1997$
G	6187	$p_G \approx 0.2080$
T	9143	$p_T \approx 0.3073$
Total	$N = 29751$	1

## Technicalities:

- The proportions seem to add up to 1.0001 due to rounding errors, but add up to 1 if computed exactly.
- SARS is an RNA virus, so it uses U's instead of T's in RNA form. When it integrates into the host genome, it becomes DNA with T's. This is the form in which it was sequenced.

# Example: SARS — Applying the test

- For each letter, here are the  $P$ -values, using  $P = 1 - (1 - p^y)^{(1-p)^N}$  where  $N = 29751$ ;  
 $p = p_A, p_C, p_G, p_T$  is from the previous slide;  
and  $y = y_{\max}$  is determined in the data.

	Nucleotide			
	A	C	G	T
$p$	0.2851	0.1997	0.2080	0.3073
$y_{\max}$	24	6	6	7
$P$ -value	$6.1870 \cdot 10^{-9}$	0.9995	0.9999	1.0000

- For A at significance level  $\alpha = 0.05$ :  $P = 6.1870 \cdot 10^{-9} \leq 0.05$ .  
So  $P \leq \alpha$  and the result is significant.  
We reject the null hypothesis and accept the alternative.
- For long runs of C, G, or T:  
the  $P$ -value is greater than 0.05 so the result is not significant.  
We accept the null hypothesis in each of those cases.

# Homopolymers (repeats of one letter)

- It turns out the long run of A's is the final 24 letters of the genome sequence (a “poly(A) tail”):
  - If we omit those,  $p_A$  goes down to
$$p_A = (8481 - 24)/(29751 - 24) = 8457/29727 = 0.2845,$$
and the next longest run of A's has length 8.
  - This gives a  $P$ -value  $P = 0.5985$ .
  - Since  $P > \alpha$  ( $0.5985 > .05$ ) the result is not significant.
- Poly(A) tails of up to several hundred A's occur at the 3' end of mRNA in eukaryotic mRNA.
- Once there are a few of the same nucleotide in a row, it is thought that DNA polymerase suffers from “slippage” and the number of repetitions lengthens over evolutionary time.
- 454 sequencing is error-prone in homopolymeric regions. It adds as many of the same nucleotide as possible in one cycle, stained with a dye, but the light output isn't proportional to the number of nucleotides incorporated.

# Computing the $P$ -value, other methods

- The book has three estimates of the  $P$ -value.
- $P_1 = 1 - (1 - p^y)^n = 1 - (1 - p^y)^{(1-p)N}$  (the one we did).
- When  $N$  is large and  $(1 - p)Np^y \leq 1$ , this is approximately  $P_2 = 1 - e^{-(1-p)Np^y}$ .
- $P_3$  treats  $n$  as a random variable, with  $n \sim \text{Binomial}(N, 1 - p)$ :  
$$P(n = k) = \binom{N}{k} (1 - p)^k p^{N-k} \quad \text{for } k = 0, 1, \dots, N$$

$$\begin{aligned} P_3 &= P(Y_{\max} \geq y) \\ &= \sum_{k=0}^N P(n = k) P(Y_{\max} \geq y | n = k) \\ &= \sum_{k=0}^N \binom{N}{k} (1 - p)^k p^{N-k} \cdot (1 - (1 - p^y)^k) \\ &= \sum_{k=0}^N \binom{N}{k} (1 - p)^k p^{N-k} - \sum_{k=0}^N \binom{N}{k} ((1 - p)(1 - p^y))^k p^{N-k} \\ &= ((1 - p) + p)^N - ((1 - p)(1 - p^y) + p)^N \\ &= 1^N - (1 - (1 - p)p^y)^N = 1 - (1 - (1 - p)p^y)^N \end{aligned}$$

# Table of $P$ -values

The table below is the  $P$ -values computed all three ways for the longest repeat of  $A$ 's with  $p = p_A = .25$  and sequence length  $N = 100000$ .

$y$	$P_1 = 1 - (1 - p^y)^{(1-p)N}$	$P_2 = 1 - e^{-(1-p)Np^y}$	$P_3 = 1 - (1 - (1-p)p^y)^N$
$\leq 5$	1.	1.	1.
6	0.99999999889	0.99999999888	0.99999999889
7	0.9897223095	0.9897208398	0.9897219505
8	0.6815910548	0.6815880136	0.6815903598
9	0.2488147944	0.2488142305	0.2488128140
10	0.0690293562	0.0690275311	0.0690316757
11	0.0177211033	0.0177224700	0.0177211028
12	0.0044600246	0.0044603712	0.0044600245
13	0.0011168758	0.0011169628	0.0011193730
14	0.0002774615	0.0002793577	0.0002799608
15	0.0000674977	0.0000698468	0.0000699976

Taylor series can be used to show why these are very close.

$$P_1 = 1 - (1 - u)^{Nv}, P_2 = 1 - e^{Nuv}, P_3 = 1 - (1 - uv)^N$$

with  $u = p^y$  and  $v = 1 - p$  and  $Nu \ll 1$ .

# Errors in hypothesis testing

## Terminology: Type I or II error

Decision	True state of nature	
	$H_0$ true	$H_1$ true
Accept $H_0$ / Reject $H_1$	Correct decision	Type II error
Reject $H_0$ / Accept $H_1$	Type I error	Correct decision

## Alternate terminology:

Null hypothesis  $H_0$  = “negative”  
Alternative hypothesis  $H_1$  = “positive”

Decision	True state of nature	
	$H_0$ true	$H_1$ true
Acc. $H_0$ / Rej. $H_1$ / “negative”	True Negative (TN)	False Negative (FN)
Rej. $H_0$ / Acc. $H_1$ / “positive”	False Positive (FP)	True Positive (TP)

# Measuring $\alpha$ and $\beta$ from empirical data

Suppose you know the # times the tests fall in each category

Decision	True state of nature		Total
	$H_0$ true	$H_1$ true	
Accept $H_0$ / Reject $H_1$	1	2	3
Reject $H_0$ / Accept $H_1$	4	10	14
Total	5	12	17

## *Error rates*

**Type I error rate:**  $\alpha = P(\text{reject } H_0 | H_0 \text{ true}) = 4/5 = .8$

**Type II error rate:**  $\beta = P(\text{accept } H_0 | H_0 \text{ false}) = 2/12 = 1/6$

## *Correct decision rates*

**Specificity:**  $1 - \alpha = P(\text{accept } H_0 | H_0 \text{ true}) = 1/5 = .2$

**Sensitivity:**  $1 - \beta = P(\text{reject } H_0 | H_0 \text{ false}) = 10/12 = 5/6$

Power = sensitivity =  $5/6$