

## **BST 140.652 Review notes**

1. You are responsible for the correctness of all of the formulae on this review sheet. (There are undoubtedly typographical errors :-).
2. You should know, *and understand*, everything in these review notes.
3. You can bring a *non-fancy* (you know what I mean) scientific calculator. It must be able to take logs and raise numbers to exponents.
4. Do not bring in a calculator that beeps every time a button is pushed.
5. You can bring in one sheet of  $8.5 \times 11$  paper filled, front and back, with formulae and notes.

# 1 Comparing two binomials

1. Let  $X \sim \text{Binomial}(n_1, p_1)$  and  $\hat{p}_1 = X/n_1$
2. Let  $Y \sim \text{Binomial}(n_2, p_2)$  and  $\hat{p}_2 = Y/n_2$
3. We also use the following notation:

$n_{11} = X$	$n_{12} = n_1 - X$	$n_{1+} = n_1$
$n_{21} = Y$	$n_{22} = n_2 - Y$	$n_{2+} = n_2$
$n_{+1}$	$n_{+2}$	

4. Consider testing  $H_0 : p_1 = p_2$  versus  $H_1 : p_1 \neq p_2$ ,  $H_2 : p_1 > p_2$ ,  $H_3 : p_1 < p_2$
5. A useful statistic compares the differences in proportions

$$TS = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where  $\hat{p} = \frac{X+Y}{n_1+n_2}$  is the estimate of the common proportion under the null hypothesis. This statistic is normally distributed for large  $n_1$  and  $n_2$ .

6. To estimate  $p_1 - p_2$  we can use  $\hat{p}_1 - \hat{p}_2$ , which has an estimated standard error  $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ , and construct a Wald confidence interval.
7. An easy fix to improve the performance of the Wald interval is to use  $\tilde{p}_1 = (X + 1)/(n_1 + 2)$  and  $\tilde{p}_2 = (Y + 1)/(n_2 + 2)$  instead of  $\hat{p}_1$  and  $\hat{p}_2$ .
8. The **relative risk** is defined as  $p_1/p_2$  with estimate  $\hat{p}_1/\hat{p}_2$ .
9. The standard error for the *log relative risk* is

$$SE_{\log RR} = \sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}}$$

- a.  $\frac{\log \hat{RR} - \log RR}{SE_{\log RR}}$  is normally distributed for large  $n_1$  and  $n_2$
  - b. For hypothesis testing, use the null estimate of  $p$
  - c. For intervals, use  $\hat{p}_1$  and  $\hat{p}_2$  in  $\hat{SE}_{\log RR}$ . Exponentiate the interval to get one for the RR
10. The **odds ratio** is defined as  $OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$
  11. An estimate of the odds ratio is  $\hat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$
  12. A standard error for the odds ratio is  $\hat{SE}_{\log OR} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$

13. For large sample sizes  $\frac{\log \hat{OR} - \log OR}{\hat{SE}_{\log OR}}$  follows a standard normal distribution. You can use this to get a Wald confidence interval and perform hypothesis test for the OR.
14. Exponentiate to get a CI for the odds ratio.
15. The odds ratio is invariant to transposing rows and columns
16. Taking logs for the RR and OR is done b/c it their finite sample distributions are often quite skewed and convergence to normality is faster on the log scale.

## 2 The delta method

1. The **delta method** is a useful tool for obtaining asymptotic standard errors.
2. The delta method states the following. If

$$\frac{\hat{\theta} - \theta}{\hat{SE}_{\hat{\theta}}} \rightarrow N(0, 1)$$

then

$$\frac{f(\hat{\theta}) - f(\theta)}{f'(\hat{\theta})\hat{SE}_{\hat{\theta}}} \rightarrow N(0, 1).$$

3. The delta method is motivated by noting that when  $\hat{\theta}$  is close to  $\theta$  then

$$\frac{f(\hat{\theta}) - f(\theta)}{\hat{\theta} - \theta} \approx f'(\hat{\theta})$$

so that

$$\frac{f(\hat{\theta}) - f(\theta)}{f'(\hat{\theta})\hat{SE}_{\hat{\theta}}} \approx \frac{\hat{\theta} - \theta}{\hat{SE}_{\hat{\theta}}}.$$

4. Therefore the asymptotic standard error for  $f(\hat{\theta})$  is  $f'(\hat{\theta})\hat{SE}_{\hat{\theta}}$ .

## 3 Chi squared testing

1. Use the notation from Section 1.
2. The chi-squared statistic is written as

$$\sum \frac{(O - E)^2}{E}$$

the sum is taken over all four cells. The **expected** cell counts are calculated under the null hypothesis.

3. An easy computational form for this statistic is

$$\frac{n(n_{11}n_{22} - n_{12}n_{21})}{n_{1+}n_{2+}n_{+1}n_{+2}}.$$

4. We reject  $H_0 : p_1 = p_2$  if the statistic is large. It is a two sided test. Compare to a .95<sup>th</sup> quantile of the Chi-squared distribution with 1 degree of freedom.
5. The chi-squared statistic is the square of the difference in proportions statistic with the common  $p$  in the denominator.
6. The chi-squared statistic is invariant to transposing the rows and columns.
7. The chi-squared statistic also applies if the sampling is **multinomial** instead of binomial. That is if only the total sample size is fixed (and hence none of the margins).
8. In the multinomial case, the null hypothesis is that the row and column classifications are **independent**.
9. It also applies the case that the chi-squared statistic applies if the  $n_{ij}$  are so called Poisson random variables, in which the total sample size isn't constrained either.

## 4 Fisher's exact test

1. Use the notation from Section 1.
2. Fisher's exact test is "exact" because it guarantees the  $\alpha$  rate, regardless of the sample size
3. Under the null hypothesis, the distribution of  $X | X + Y = z$  is the so called **hypergeometric** distribution. The PMF for the hypergeometric distribution is

$$P(X = x | X + Y = z) = \frac{\binom{n_1}{x} \binom{n_2}{z-x}}{\binom{n_1+n_2}{z}}.$$

The possible values for  $x$  are  $\max(0, z + n_1 - n) \leq x \leq \min(z, n_1)$ .

4. This distribution can be simulated by taking  $n_{1+}$  red balls and  $n_{2+}$  white balls and randomly allocating to two bins that can hold  $n_{+1}$  and  $n_{+2}$  balls respectively.
5. For a one sided hypothesis, you can perform Fisher's exact test by calculating the hypergeometric probabilities for all tables that are as or more supportive of the alternative hypothesis. Remember to constrain the margins. To obtain the two sided P-value, double the smaller of the one sided P-values.

6. Like the chi-squared test, Fisher's exact test applies to binomial, multinomial or Poisson sampling.
7. **Exact unconditional** P-values are competitors to Fisher's exact (conditional) test. An example unconditional P-value would proceed as follows. Under the null hypothesis  $H_0 : p_1 = p_2 = p$ , (conditional) for a specific value of  $p$ , calculate the binomial probability that

$$P(TS \geq TS_{obs} | p)$$

where  $TS$  is some test statistic. An exact unconditional P-value takes the largest of these P-values (maxing over  $p$ ).

## 5 Chi-squared testing

1. The chi-squared test can be used to test  $p_1 = p_2 = \dots = p_k$  for  $k$  binomial observations,  $X_i \sim \text{Binomial}(n_i, p_i)$ .
2. The test statistic is  $\sum \frac{(O-E)^2}{E}$  where  $O$  are the observed counts (successes and failures) and  $E$  are the estimated expected counts under the null hypothesis. This statistic is a chi-square with  $k - 1$  degrees of freedom.
3. A followup test would compare the proportions individually, two at a time.
4. The test can be generalized to multicategory settings where we would want to test whether or not the distribution of the counts in each row are the same. This test would have  $(rows - 1)(cols - 1)$  degrees of freedom.
5. For multinomial sampling (only the overall sample size is constrained) a test of independence of the row and column classifications can be done. If  $n_{ij}$  are the observed counts in cell  $i, j$ , then the expected counts are  $n_{i+}n_{+j}/n$ . (Here  $n_{i+}$  refers to the  $i^{th}$  row total and  $n_{+j}$  refers to the  $j^{th}$  column total). The resulting statistic has degrees of freedom  $(rows - 1)(cols - 1)$ .
6. The test statistic for independence and the test for equal distributions in each row are mathematically the same and follow a chi-squared distribution with  $(rows - 1)(cols - 1)$  degrees of freedom. The only difference is in the interpretation of the test.
7. Exact tests of independence (generalizations of Fisher's exact test) can be performed using Monte Carlo simulation.
8. Goodness of fit testing tests whether or not a series of counts follow a specified distribution. That is  $H_0 : p_1 = p_{01}, p_2 = p_{02}, \dots, p_k = p_{0k}$  where  $p_{0i}$  are specified. The expected count for cell  $i$  is  $n * p_{0i}$ . The resulting statistic has  $k - 1$  degrees of freedom.

## 6 Retrospective case/control studies

1. In **retrospective** case/control studies, case status is fixed while disease (or exposure) status is allowed to vary.
2. In such a setting  $P(\text{Case} \mid \text{Exposure})$  is not estimable. However,  $P(\text{Exposure} \mid \text{Case})$  is.
3. The both the sample and population odds ratios are invariant to transposing rows and columns. As a consequence, the odds ratio comparing case status given exposure is equal to the odds ratio comparing exposure status given being a case. That is

$$\frac{\text{Odds}(\text{case} \mid \text{exposure})}{\text{Odds}(\text{case} \mid \text{exposure}^c)} = \frac{\text{Odds}(\text{exposure} \mid \text{case})}{\text{Odds}(\text{exposure} \mid \text{case}^c)}$$

Here a superscript  $c$  denotes the compliment. Therefore the prospective odds ratio of interest is in fact estimable given retrospective sampling.

4. The  $OR$  is related to the  $RR$  by

$$OR = RR \times \frac{1 - P(\text{case} \mid \text{exposure}^c)}{1 - P(\text{case} \mid \text{exposure})}$$

Therefore, the  $OR$  will be larger than the  $RR$  if the disease is more prevalent amongst the exposed and will be smaller if the disease is more prevalent amongst the unexposed. If the disease is rare in both the exposed and unexposed groups, the  $OR$  approximates the relative risk.

## 7 Exact inference for the odds ratio

1. The logit is the **log of the odds** ( $\text{logit}(p) = \log\{p/(1-p)\}$ ).
2. Differences in logits are **log odds ratios**.
3. If  $\eta = \log\{p/(1-p)\}$  is the logit of  $p$  then **inverse logit** function is  $p = e^\eta/(1 + e^\eta)$ .
4. Assume  $X$  Binomial( $n_1, p_1$ ) and  $Y$  Binomial( $n_2, p_2$ ) where
  - a.  $\text{logit}(p_1) = \delta$  hence  $p_1 = e^\delta/(1 + e^\delta)$ .
  - b.  $\text{logit}(p_2) = \delta + \theta$  hence  $p_2 = e^{\delta+\theta}/(1 + e^{\delta+\theta})$ .
5. Then  $\theta$  is the log odds ratio and  $e^\theta$  is the odds ratio
6. Then  $\delta$  is a nuisance parameter.

7. We can eliminate  $\delta$  from consideration by considering the distribution of  $X$  given  $X + Y$

$$P(X = x | X + Y = z; \theta) = \frac{\binom{n_1}{x} \binom{n_2}{z-x} e^{x\theta}}{\sum_u \binom{n_1}{u} \binom{n_2}{z-u} e^{u\theta}}$$

The possible values for  $x$  (and the range of the sum in the denominator) are  $\max(0, z + n_1 - n) \leq x \leq \min(z, n_1)$ . When  $\theta = 0$  this distribution is exactly the hypergeometric distribution used for Fisher's exact test.

8. The PMF  $P(X = x | X + Y = z; \theta)$  can be used to construct exact tests of  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta > \theta_0$ ,  $H_2 : \theta < \theta_0$  and  $H_3 : \theta < \theta_0$ . Notice Fisher's exact test above only considered the specific hypothesis  $\theta_0 = 0$ . By inverting these exact tests, taking those values of  $\theta_0$  for which we fail to reject, we can construct exact confidence intervals for the log odds ratio (and hence the odds ratio).
9. We can also use the conditional likelihood  $\mathcal{L}(\theta) = P(X = x | X + Y = z; \theta)$  to plot a likelihood function for  $\theta$  alone (without having to consider  $\delta$ ).
10. It is known that by conditioning on  $X + Y$  we eliminate  $\delta$ , yet we also discard some information regarding  $\theta$ . The argument over whether to condition on  $X + Y$  or to consider other approaches has lasted nearly 100 years.

## 8 Multiplicity

1. When conducting  $k$  hypothesis tests, the **familywise error rate** refers to the probability of falsely rejecting the null hypothesis in any of the  $k$  tests.
2. Bonferroni's inequality implies that the familywise error rate is no larger than  $k\alpha$  where  $\alpha$  is the Type I error rate (applied to each test individually). Therefore a **Bonferroni adjustment** uses the Type I error rate  $\alpha^* = \alpha/k$  for each test. Under this adjustment the familywise error rate is no larger than  $\alpha$ .
3. If there are a large number tests whose outcomes are independent (which is rarely the case), then the Bonferroni bound on the family wise error rate is nearly attained.
4. A lower bound on the family wise error rate is the common Type I error rate used for each test. This lower bound is attained if the outcome of the tests are perfectly dependent. Therefore, when the test outcomes are correlated the Bonferroni correction, which uses  $\alpha/k$  as the common error rate, can be very conservative.
5. The **false discovery rate** is defined as the proportion of tests that are falsely declared significant.
6. The Benjamini and Hochberg procedure to control the FDR follows as

- i. Order your p-values so that  $p_1 < \dots < p_k$
  - ii. Define  $q_i = kp_i/i$
  - iii. Define  $F_i = \min(q_i, \dots, q_k)$
  - iv. Reject  $H_0$  for all  $i$  so that  $F_i$  is less than the desired FDR. (Because the  $F_i$  are increasing, one need only find the largest  $i$  so that  $F_i < \text{FDR}$ ).
7. Both FDRs and Bonferoni corrections are only useful insofar as protecting against errors when invoking a decision rule. The actual evidence in each test visa vis the two hypothesis is unaffected by the number of tests in question.

## 9 Stratified two-by-two tables

1. **Simpson's paradox** refers to the fact that the apparent marginal relationship between two variables can be different (and even reversed in direction) than the conditional relationship given a third variable. We reviewed an example where the probability of receiving the death penalty was higher for white defendants than black defendants marginally. However, when the race of the victim was considered, the probability of receiving the death penalty was higher for blacks than whites when the victims were either white or black.
2. Simpson's paradox illustrates that variables that are correlated with both the explanatory and outcome variable of interest can distort the estimated effect. One strategy to address such **confounding** is to **stratify** by the confounding variable and combine the strata-specific estimates. Stratifying does not come for free, however, as unnecessary stratification can reduce the precision of estimates.
3. Let  $n_{ijk}$  be the  $(i, j)^{th}$  counts from the  $k^{th}$  table for a series of  $2 \times 2$  tables.
4. Let  $\hat{\theta}_k = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}$  be the  $k^{th}$  sample odds ratio.
5. Let  $r_k = \frac{n_{12k}n_{21k}}{n_{++k}}$
6. Let  $s_k = \frac{n_{11k}n_{22k}}{n_{++k}}$
7. The **Mantel Haenszel** estimate of the common odds ratio across  $k$  strata is

$$\hat{\theta}_{MH} = \frac{\sum_k r_k \hat{\theta}_k}{\sum_k r_k} = \frac{\sum_k s_k}{\sum_k r_k} = \frac{\sum_k n_{11k}n_{22k}/n_{++k}}{\sum_k n_{12k}n_{21k}/n_{++k}}.$$

If the odds ratios are constant across the strata, then the MH estimates this common odds ratio. If not and the strata can be thought of as a random draw from a population, the MH estimate is an estimate of the population averaged OR (averaged across the stratifying variables).

8. The standard error for the log of the MH estimate is

$$\begin{aligned} & \frac{1}{2(\sum_k s_k)^2} \sum_k n_{++k}^{-1} (n_{11k} + n_{22k}) s_k \\ & + \frac{1}{2(\sum_k r_k)^2} \sum_k n_{++k}^{-1} (n_{12k} + n_{21k}) r_k \\ & + \frac{1}{2(\sum_k r_k)(\sum_k s_k)} \sum_k n_{++k}^{-1} [(n_{11k} + n_{22k}) s_k + (n_{12k} + n_{21k}) r_k]. \end{aligned}$$

9. The **Cochran, Mantel, Haenszel** (CMH) test considers the hypothesis  $H_0 : \theta_1 = \dots = \theta_k = 1$ . This null hypothesis is equivalent to specifying that the response and exposure variables are independent given the stratifying variable. The test is most powerful for alternatives of the form  $H_a : \theta_1 = \dots = \theta_k \neq 1$ .

10. The CMH test conditions on the rows and columns of each of the  $2 \times 2$  contingency tables. Under this conditioning

a.  $E(n_{11k}) = n_{1+k}n_{+1k}/n_{++k}$

b.  $\text{Var}(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^2(n_{++k} - 1)$

11. The CMH test statistic is

$$\frac{[\sum_k \{n_{11k} - E(n_{11k})\}]^2}{\sum_k \text{Var}(n_{11k})}.$$

This statistic follows a  $\chi^2(1)$  distribution for large sample sizes.

12. It is possible to perform an exact test for small sample sizes using the hypergeometric distributions from each table.

13. It is often a good idea to precede a CMH test with a test for heterogeneity of the odds ratios. This test for the possibility of **effect modification**; that is, the effect of the exposure on the response differs at different levels of the stratifying variable.

14. Using so called random effect logit models, one can perform CMH-like analyses using a model based approach.

## 10 Models for matched pairs binary data

1. In this section we are concerned with comparing dependent proportions. Dependence can arise from **repeated sampling** or **matching**, for example. A distinguishing feature of this section is that the same response is measured twice (either at two occasions or via matching).

2. Consider taking a presidential approval poll amongst registered voters at two times: say November 2002 and November 2005. Consider the distinction between

- a. polling an independent group each time
- b. polling the same people twice

In the latter case the binary observations are **dependent**. Sampling according to a. allows one to measure any **cross-sectional** changes in approval. Sampling according to b. allows one to measure any **longitudinal changes** in approval. These two are not the same. Imagine if every individual's approval status remained constant yet half of the people who did not approve renounced their citizenship and moved to Canada (hence were no longer eligible for the poll). The longitudinal effect is constant (there was no change in individual's opinions), while the cross-sectional approval rating will have gone up dramatically.

3. The methods described also apply to matched case control studies. Here the response is diseased or not and the dependence is due to the fact that the cases and controls were matched
4. Let  $n_{ij}$  be the cell counts in the  $2 \times 2$  table.  $i = 1, 2, j = 1, 2$

	time 2		
time 1	Yes	No	Total
Yes	$n_{11}$	$n_{12}$	$n_{1+}$
no	$n_{21}$	$n_{22}$	$n_{2+}$
Total	$n_{+1}$	$n_{+2}$	$n$

	Controls		
Cases	$D$	$\bar{D}$	Total
$D$	$n_{11}$	$n_{12}$	$n_{1+}$
$\bar{D}$	$n_{21}$	$n_{22}$	$n_{2+}$
Total	$n_{+1}$	$n_{+2}$	$n$

Notice that the case control data is organized differently than before.

5. Let  $\pi_{ij}$  be the cell probabilities.
6. Consider testing  $H_0 : \pi_{1+} = \pi_{+1}$  versus  $H_0 : \pi_{1+} \neq \pi_{+1}$ , **marginal homogeneity**. Notice that this hypothesis is the same as  $H_0 : \pi_{12} = \pi_{21}$ .
7. **McNemar's** test statistic is

$$\frac{(n_{21} - n_{12})^2}{n_{21} + n_{12}}$$

which follows a chi-squared distribution with 1 degree of freedom. Notice that only the discordant cells enter in to this test statistic.

8. It turns out that under the null hypothesis

$$P(n_{12} = x \mid n_{12} + n_{21} = z) = \binom{z}{x} \left(\frac{1}{2}\right)^z.$$

That is the distribution of  $n_{12}$  given  $n_{12} + n_{21}$  is binomial with  $n_{12} + n_{21}$  trials and success probability  $1/2$ . This distribution can be used to construct exact p-values for  $H_0 : \pi_{12} = \pi_{21}$  versus  $H_1 : \pi_{12} > \pi_{21}$  or  $H_2 : \pi_{12} < \pi_{21}$ . To obtain the two sided hypothesis, double the minimum of the two one sided

9. The obvious estimator of the difference in the marginal probabilities,  $\pi_{1+} - \pi_{+1}$ , is  $d = n_{1+}/n - n_{+1}/n$ . The variance of this estimator is

$$\sigma_d^2 = \{\pi_{1+}(1 - \pi_{1+}) + \pi_{+1}(1 - \pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})\}/n$$

$\frac{d - (\pi_{1+} - \pi_{+1})}{\sigma_d}$  follows an asymptotic normal distribution. Notice that  $\sigma_d^2$  is nearly identical to the statistic obtained with independent proportions. When would the standard error be lower?

10. The odds ratio comparing the marginal probabilities is

$$\frac{\pi_{1+}/\pi_{2+}}{\pi_{+1}/\pi_{+2}} = \frac{\pi_{1+}\pi_{+2}}{\pi_{+1}\pi_{2+}}$$

11. The maximum likelihood estimate of the marginal log odds ratio is  $\hat{\theta} = \log\{\hat{\pi}_{1+}\hat{\pi}_{+2}/\hat{\pi}_{+1}\hat{\pi}_{2+}\}$  where  $\hat{\pi}_{ij} = n_{ij}/n$  are the sample proportions. The asymptotic variance of this estimator is

$$\{(\pi_{1+}\pi_{2+})^{-1} + (\pi_{+1}\pi_{+2})^{-1} - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2})\}/n.$$

12. Connection with the CMH test. Each subject's (or matched pair's) responses can be represented as one of four tables.

	Response			Response	
Case Status	$D$	$\bar{D}$	Case Status	$D$	$\bar{D}$
Case	1	0	Case	1	0
Control	1	0	Control	0	1
	Response			Response	
Case Status	$D$	$\bar{D}$	Case Status	$D$	$\bar{D}$
Case	0	1	Case	0	1
Control	1	0	Control	0	1

McNemar's test statistic is equivalent to the CMH test where subject is the stratifying variable and each  $2 \times 2$  table is the observed zero-one table for that subject.

## 11 Poisson

1. The **Poisson distribution** is used to model counts. The mass function is

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

for  $x = 0, 1, \dots$

2. The mean of a Poisson random variable is  $\lambda$
3. The variance of a Poisson random variable is  $\lambda$
4. The Poisson distribution arises in several ways. The classic construction uses Poisson processes. A **Poisson process** counts the number of events occurring in a particular unit of time. Think of the number of people waiting for the bus in one hour. Consider decomposing the time unit into very small components, say of length  $h$ . Then the number of events per time unit will be Poisson with mean  $\lambda$  if
  - The probability of an event in an interval of length  $h$  is approximately  $\lambda h$  while the probability of more than one event is very small (and goes to 0 as  $h$  gets smaller and smaller).
  - Whether or not an event occurs in one interval is independent of whether or not an event occurs in another.
5. If the number of events from a Poisson process evaluated for 1 time unit is Poisson( $\lambda$ ) then the number of events considering  $t$  time units is Poisson( $t\lambda$ ). For example, if the number of people waiting for a bus in one hour is Poisson with  $\lambda = 10$  people per hour then the number of people waiting for a bus in two hours is Poisson with  $\lambda = 20$ .
6. Another way to motivate the Poisson distribution is that it is an approximation to the binomial distribution when  $n$  is large and  $p$  is small. In this case  $\lambda = np$ .
7. If  $X_i$  are Poisson( $\lambda t_i$ ) then

$$\sum_i X_i \sim \text{Poisson}(\lambda \sum_i t_i).$$

Also, all of the relevant information regarding  $\lambda$  is stored in the sum.

8. Poisson random variables have a nifty CLT. If  $X$  is Poisson( $t\lambda$ ) then

$$\frac{X - t\lambda}{\sqrt{t\lambda}} = \frac{X - \text{Mean}}{\text{SD}}$$

converges to a standard normal as  $t\lambda$  gets large

9. If  $X \sim \text{Poisson}(t\lambda)$  then  $X/t$  is the ML estimate of  $\lambda$ .
10. An asymptotic test of  $H_0 : \lambda = \lambda_0$  (versus any of the three alternatives) can be performed with the Z test statistic

$$TS = \frac{X - t\lambda_0}{\sqrt{t\lambda_0}}.$$

( $t\lambda_0$  has to be large for this to work).

11. An exact test can be performed by using the Poisson distribution exactly. For example for the alternative  $H_0 : \lambda > \lambda_0$ , the exact P-value is

$$P(X \geq x_{obs} | t\lambda_0) = \sum_{x=x_{obs}}^{\infty} \frac{(t\lambda_0)^x e^{-t\lambda_0}}{x!}.$$

To get an exact P-value for a two-sided alternative, double the smaller of the one sided P-values.

12. If  $X_1 \sim \text{Poisson}(t_1\lambda_1)$  and  $X_2 \sim \text{Poisson}(t_2\lambda_2)$  then consider testing  $H_0 : \lambda_1 = \lambda_2 = \lambda$  versus  $H_a : \lambda_1 \neq \lambda_2$ .

- a. Under  $H_0$  the estimated expected count in Group 1 is

$$E_1 = \hat{\lambda}t_1 = (x_1 + x_2) \frac{t_1}{t_1 + t_2}$$

- b. For Group 2

$$E_2 = \hat{\lambda}t_2 = (x_1 + x_2) \frac{t_2}{t_1 + t_2}$$

The test stastic

$$TS = \sum \frac{(O - E)^2}{E} = \frac{(x_1 - E_1)^2}{E_1} + \frac{(x_2 - E_2)^2}{E_2}$$

follows a Chi-squared distribution with 1 df, with large values supporting the alternative.

13. The test from the previous item is equal to

$$TS = \frac{(X_1 - E_1)^2}{V_1}$$

where

$$V_1 = (x_1 + x_2)t_1t_2/(t_1 + t_2)^2$$

14. For one sided alternative, compare the test stat  $TS = \frac{X_1 - E_1}{\sqrt{V_1}}$  to a standard normal.
15. The **relative rate** is  $\lambda_1/\lambda_2$  which has the obvious estimate  $(x_1/t_1)/(x_2/t_2)$ . The standard error for the **log relative rate** is

$$\sqrt{\frac{1}{x_1} + \frac{1}{x_2}}$$

16. The Poisson model for time-to-event data is **likelihood equivalent** to a model of independent exponentials. In particular, let  $n$  individuals be followed over time so that  $x$  experience events (say death) while  $n - x$  were **censored** (the end of the study occurred without an event for these people). If we model a person's time to event,  $Y_i$ , as  $\text{Exponential}(\lambda)$  then:

- a. The contribution to the likelihood for the non-censored individuals is

$$\lambda e^{-y_i \lambda}$$

- b. The contribution to the likelihood for the censored individuals is

$$P(Y \geq y_i; \lambda) = \int_{y_i}^{\infty} \lambda e^{-u \lambda} du = e^{-y_i \lambda}$$

- c. So the likelihood is

$$\left\{ \prod_{\text{non-censored } i} \lambda e^{-y_i \lambda} \right\} \times \left\{ \prod_{\text{censored } i} e^{-y_i \lambda} \right\} = \lambda^n \exp \left( -\lambda \sum_{i=1}^n y_i \right)$$

This is the same likelihood as one would obtain if  $X \sim \text{Poisson}(\lambda \sum y_i)$ .

## 12 Non-parametric testing

1. Non-parametric testing relaxes the assumptions of parametric tests. There are also referred to as “distribution free” tests. Note that these tests are not “assumption free”.
2. For paired continuous data, consider taking the differences (as in the paired T-test); denote these differences by  $D_i$ . If the median difference is 0, then  $p = P(D_i > 0) = .5$ ; if the median difference is greater than 0, then  $P(D_i > 0) > .5$ , and so on. The **sign test** tests  $H_0 : p = .5$  versus the three alternative using the indicators of whether each  $D_i$  is larger than 0. Let  $D_+$  be the total number of positive differences. Then  $D_+$  is Binomial with success probability  $p$ . All of the usual binomial procedures can then be used to carry out the tests. Instances where  $D_i = 0$  are thrown out and the overall sample size reduced.
3. The sign test disregards a lot of information contained in the observations. The **signed rank test** overcomes this to a large degree by also incorporating the **ranks** of the observations. The signed rank procedure is as follows
  - a. Take the paired differences
  - b. Take the absolute values of the differences
  - c. Rank these absolute values, throwing out the 0s
  - d. Multiply the ranks by the sign of the difference (+1 for a positive difference and -1 for a negative difference)
  - e. Calculate the rank sum  $W_+$  of the positive ranks
4. For small sample sizes,  $W_+$  has an exact distribution under the null hypothesis. Critical values can be obtained from tables. If the alternative is that the median difference is larger than 0, then  $W_+$  should be large (hence reject if it is larger than the critical value). Vice-versa for the median difference being smaller than 0. Also most books will provide the relevant critical values for the two sided test.

5. A large sample test statistic can be constructed as follows

$$E(W_+) = n(n + 1)/4$$

$$Var(W_+) = n(n + 1)(2n + 1)/24$$

$$TS = \{W_+ - E(W_+)\}/Sd(W_+) \rightarrow \text{Normal}(0, 1)$$

6. With ties in the ranks, a correction term can be added. Instances where  $D_i = 0$  are thrown out and the overall sample size reduced.

7. For unpaired data the relevant test is called the **rank sum** test.

8. Procedure

(a) Discard the treatment labels

(b) Rank the observations

(c) Calculate the sum of the ranks in the first treatment

(d) Either

\* calculate the asymptotic normal distribution of this statistic

\* compare with the exact distribution under the null hypothesis

9. Let  $W$  be the sum of the ranks for the first treatment ( $A$ )

Let  $n_A$  and  $n_B$  be the sample sizes

Then

- $E(W) = n_A(n_A + n_B + 1)/2$

- $Var(W) = n_A n_B (n_A + n_B + 1)/12$

- $TS = \{W - E(W)\}/Sd(W) \rightarrow N(0, 1)$