A9: Statistics

Sheet 3 — HT 2022

(Lectures 8–11, Notes section 3, HT21 videos 14–20)

1. The heart rate (beats per minute) of 10 children was measured in two situations: (i) at rest, and (ii) in anticipation of them doing a minute's exercise. The data are given below.

Rest, x	72	116	79	97	90	67	115	82	95	82
Anticipation, y	76	120	84	99	93	75	116	83	98	87

The sample means and variances are $\overline{x} = 89.5$, $s_x^2 = 274.9$, $\overline{y} = 93.1$, $s_y^2 = 238.8$.

(a) Assuming the data are normally distributed, carry out a two-sample t-test of the null hypothesis that the mean heart rate for the two situations is the same, against the alternative that it is different. What further assumptions are required for the test to be valid?

How would you modify the test if the alternative is that the mean heart rate is *higher* in situation (ii)? Explain which alternative you think is more appropriate here.

(b) Suggest a more appropriate test than that in (a). Carry out this test and explain why you prefer it.

2. Let $X_1, \ldots, X_n$ be independent $N(\theta, \sigma_0^2)$ random variables, where σ_0^2 is known. Find the most powerful test of size α of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, where $\theta_1 > \theta_0$.

Show that the power function $w(\theta)$ of this test is given by

$$w(\theta) = 1 - \Phi\left(\frac{\sqrt{n}}{\sigma_0}(\theta_0 - \theta) + z_\alpha\right)$$

where Φ is the cumulative distribution function of the standard normal distribution and $\Phi(z_\alpha) = 1 - \alpha$.

If $\theta_0 = 0$, $\theta_1 = 0.5$ and $\sigma_0 = 1$, how large must n be if $\alpha = 0.05$ and the power at θ_1 is to be 0.975? [If Φ is the $N(0,1)$ cdf, then $\Phi(1.645) = 0.95$ and $\Phi(1.96) = 0.975$.]

Neil Laws: neil.laws@stats.ox.ac.uk

3. A telephone receptionist for a large partnership of financial advisers is responsible for determining the precise nature of each incoming enquiry and connecting the client with an appropriate adviser. The number of inappropriate connections on any given day may be modelled by a random variable X which has a Poisson distribution with mean μ. If Z is the number of inappropriate connections made over a period of n days, determine the distribution of Z and find its expected value.

   Uhura, who has been such a receptionist for many years, has been found to have a mean rate of $\mu_U = 0.47$ inappropriate connections per day. For several months she has been training Spock, a new receptionist, with corresponding mean rate μ_S. At a meeting of senior partners, it was conjectured that Spock was already as proficient as Uhura; accordingly they resolved to keep a daily record of the number of inappropriate connections made by him over his next 10 working days. Find a critical region of size 5% for a test of the hypothesis that Spock is as proficient as Uhura versus the alternative that he is less proficient.

   For what values of μ_S does the probability of type II error fall below 10%?

   [Note that if $\varphi_\mu(k) = \sum_{x=0}^{k} \mu^x e^{-\mu}/x!$, then $\varphi_{4.7}(8) = 0.95$, $\varphi_{13}(8) = 0.1$.]

4. When studying the sex ratio in a population using a sample of size n, it is usually assumed that, indpendently, each child is male with probability p. Renkonen (1956) observed 19,711 male births out of a total of 38,562 births in American families with two children each. Use the likelihood ratio statistic Λ to test the hypothesis $H_0 : p = \frac{1}{2}$ against a suitable alternative which you should specify.

   Renkonen also found 17,703 males out of 35,042 similar births in Finland. Use the generalised likelihood ratio test to test the hypothesis that p has the same value in each country versus a suitable alternative.

Neil Laws: neil.laws@stats.ox.ac.uk

5. (a) A random variable X has a distribution given by

$$P(X = i) = \pi_i, \quad i = 1, \ldots, k$$

where $\sum_{i=1}^{k} \pi_i = 1$. In a sample of size n from a population with distribution X, the frequency of outcome i is n_i, where $n_i > 0$ and $\sum_{i=1}^{k} n_i = n$. Find the maximum likelihood estimates of $\pi_1, \ldots, \pi_k$.

(b) The leaves of the plant *Pharbitis nil* can be variegated or unvariegated and, at the same time, faded or unfaded. In an experiment reported by Bailey (1961), of 290 plants which were observed, 31 had variegated faded leaves, 37 had variegated unfaded leaves, 35 had unvariegated faded leaves and 187 had unvariegated unfaded leaves.

If the properties of variegated appearance and faded appearance are assumed independent, then a model for the above observations has respective probabilities $\frac{1}{16}, \frac{3}{16}, \frac{3}{16}, \frac{9}{16}$. The general alternative is that the probabilities π_i, $i = 1, \ldots, 4$, are restricted only by the constraint $\sum \pi_i = 1$. Use a χ^2 goodness-of-fit test to show that the data offer strong evidence that the independence model is inappropriate.

(c) A genetic theory which allows for an effect called *genetic linkage* assumes a probability model for the above observations with respective probabilities

$$\frac{1}{16} + \theta, \ \frac{3}{16} - \theta, \ \frac{3}{16} - \theta, \ \frac{9}{16} + \theta.$$

Find the equation satisfied by the maximum likelihood estimate $\widehat{\theta}$ of θ.

You may assume that $\widehat{\theta} = 0.058$.

Let H_0 be the null hypothesis that the genetic linkage model is appropriate, and let H_1 be the general alternative. If L_0 is the supremum of the likelihood under H_0 and if L_1 is the supremum of the likelihood under H_1, show that

$$\Lambda = 2 \sum_{i=1}^{4} n_i \log \left(\frac{n_i}{n \pi_i(\widehat{\theta})} \right)$$

where $\Lambda = -2(\log L_0 - \log L_1)$. Write down the approximate distribution of Λ.

What can you infer about the plausibility of the genetic linkage model?

Neil Laws: neil.laws@stats.ox.ac.uk

6. The ordered pairs of random variables (X_k, Y_k), $k = 1, \ldots, n$, are independent and

$$P\big((X_k, Y_k) = (i, j)\big) = \pi_{ij}, \quad i = 1, \ldots, r, \quad j = 1, \ldots, c$$

where $\sum_{i,j} \pi_{ij} = 1$. The frequency of the outcome (i, j) is n_{ij}, where $n_{ij} > 0$.

Find the maximum likelihood estimates of the π_{ij} assuming that:

(i) $\pi_{ij} = \alpha_i \beta_j$ for $i = 1, \ldots, r$ and $j = 1, \ldots, c$, where $\sum_i \alpha_i = \sum_j \beta_j = 1$, and

(ii) without this assumption.

Hence find test statistics for testing the null hypothesis that the X_k and the Y_k are independent using:

(a) the likelihood ratio method,

(b) Pearson's χ^2 statistic.

What can you say about the distributions of these two statistics for large values of n?

The data below (Agresti, 2007) cross-classifies gender and political party identification in the USA: 2757 individuals indicated whether they identified more strongly with the Democratic or Republican party or as Independents. Is there an association between gender and political party identification?

	Party Identification		
	Democrat	Independent	Republican
Female	762	327	468
Male	484	239	477

Neil Laws: neil.laws@stats.ox.ac.uk

7. Carry out the numerical calculations required for this sheet using R.

```
#### Question 1
x <- c(72, 116, 79, 97, 90, 67, 115, 82, 95, 82)
y <- c(76, 120, 84, 99, 93, 75, 116, 83, 98, 87)
m <- 10
n <- 10

xbar <- mean(x)
ssqx <- var(x)
ybar <- mean(y)
ssqy <- var(y)

ss <- ((m-1)*ssqx + (n-1)*ssqy) / (m+n-2)
s <- sqrt(ss)
tobs <- (xbar - ybar) / (s*sqrt(1/m + 1/n))

# since tobs is negative
2 * pt(tobs, df = 18)
pt(tobs, df = 18)

qt(0.1, df = 18)

# as a check
t.test(x, y, var.equal = TRUE)

# now paired
d <- y - x
t1 <- mean(d)/sqrt(var(d)/10)
1 - pt(t1, df = 9)
# as a check
t.test(d)

#### Question 4
x1 <- 19711
n1 <- 38562
p1hat <- x1/n1
Lambda <- 2 * ( n1*log(2) + x1*log(p1hat) + (n1-x1)*log(1-p1hat) )
1 - pchisq(Lambda, df = 1)

x2 <- 17703
```


Neil Laws: neil.laws@stats.ox.ac.uk

```
n2 <- 35042
p2hat <- x2/n2
phat <- (x1 + x2)/(n1 + n2)
term1 <- (phat/p1hat)^x1 * ((1-phat)/(1-p1hat))^(n1-x1)
term2 <- (phat/p2hat)^x2 * ((1-phat)/(1-p2hat))^(n2-x2)
ratio <- term1*term2
Lambda1 <- -2*log(ratio)
1 - pchisq(Lambda1, df = 1)

# same as Lambda1
Lambda2 <- -2 * ((x1+x2)*log(phat) + (n1+n2-x1-x2)*log(1-phat)
                 - x1*log(p1hat) - (n1-x1)*log(1-p1hat)
                 - x2*log(p2hat) - (n2-x2)*log(1-p2hat))

#### Question 5
obs <- c(31, 37, 35, 187)
expect <- 290*c(1/16, 3/16, 3/16, 9/16)
L1 <- 2 * sum(obs * log(obs/expect))
P1 <- sum((obs - expect)^2/expect)
1 - pchisq(L1, df = 3)
1 - pchisq(P1, df = 3)

n1 <- 31
n2 <- 37
n3 <- 35
n4 <- 187
a <- - 16^2*n1 - 16^2*(n2+n3) - 16^2*n4
b <- - 96*n1 - 160*(n2+n3) + 32*n4
c <- 27*n1 - 9*(n2+n3) + 3*n4
theta1 <- (-b + sqrt(b^2-4*a*c))/(2*a)
theta2 <- (-b - sqrt(b^2-4*a*c))/(2*a)

# theta1 not a valid value of theta
c(1/16+theta1, 3/16-theta1, 3/16-theta1, 9/16+theta1)

# theta2 is a valid value
c(1/16+theta2, 3/16-theta2, 3/16-theta2, 9/16+theta2)

# the log-likelihood is maximised at theta2 - picture
theta <- seq(-0.05, 0.18, length.out=50)
```

```
plot(theta, n1*log(1+16*theta) + (n2+n3)*log(3-16*theta)
     + n4*log(9+16*theta), type = "l", ylab = "g(theta)")
abline(v = theta2, lty = 2)

expect2 <- 290*c(1/16+theta2, 3/16-theta2, 3/16-theta2, 9/16+theta2)
L2 <- 2 * sum(obs * log(obs/expect2))
P2 <- sum((obs - expect2)^2/expect2)
1 - pchisq(L2, df = 2)
1 - pchisq(P2, df = 2)

#### Question 6
x <- matrix(c(762, 484, 327, 239, 468, 477), ncol = 3)
n <- sum(x)
alpha <- rowSums(x)/n
beta <- colSums(x)/n

# under the null, the expected number in cell (i,j) is n*alpha[i]*beta[j]
# an outer product denoted by %o% does exactly what we need
# e.g try
num <- 1:12
num %o% num

# so evaluate the expected numbers under the null by
expect <- n * alpha %o% beta
obs <- x

Lambda <- 2 * sum(obs * log(obs/expect))
Pearson <- sum((obs-expect)^2 / expect)
1 - pchisq(Lambda, df = 2)
1 - pchisq(Pearson, df = 2)

## as a check
chisq.test(x)
```