HW10, due April 15 2011:

8.48, 8.52, 9.14, 9.16 (pages 384 and 426)

Testing Hypotheses

Malignant Melanoma in Marathon Runners

Christina M. Ambros-Rudolph, MD; Rainer Hofmann-Wellenhof, MD; Erika Richtig, MD; Manuela Müller-Fürstner, MD; H. Peter Soyer, MD; Helmut Kerl, MD

Background: Marathon running has surged in popularity; it is generally believed to be healthy, but may be associated with medical risks. Over the past decade, we observed 8 ultramarathon runners with malignant melanoma. UV exposure, immunosuppression due to long-term intensive exercise, or both have been discussed as potential triggers in these patients. To further evaluate risk factors for malignant melanoma in marathon runners, we examined anamnestic, phenotypic, sun-related, and clinical variables in 210 athletes and compared them with those of an age- and sex-matched control group.

Observations: Although control subjects exhibited higher sun sensitivity and more common melanocytic nevi, marathon runners presented with more atypical melanocytic nevi, solar lentigines, and lesions suggestive of nonmelanoma skin cancer. These findings correlated with increasing training intensity. During exercising, most runners wore shorts (96.7%) and shirts (98.6%) that would not or would only partially cover their back and extremities. Regular use of sunscreen was reported in only 56.2% of runners.

Conclusions: Compared with a representative control group, marathon runners presented with an increased risk for malignant melanoma and nonmelanoma skin cancer. They should reduce UV exposure during exercising by choosing training and competition schedules with low sun exposure, wearing adequate clothing, and regularly using water-resistant sunscreens.

Arch Dermatol. 2006;142:1471-1474

Motivating discussion

A newspaper story discussed the possible increase of skin cancer in marathon runners, based on a paper *Malignant melanoma in marathon runners*, by Ambros-Rudolph *et al.* (a group of researchers at the Medical University of Graz, Austria).

The initial research was prompted by a group of 8 ultramarathoners with malignant melanoma (Table 1).

Table 1. Clinical and Histopathological Characteristics of 8 Ultramarathon Runners With MM*

Patient No./ Sex/Age, y	MM Location	Clark Level/Breslow Tumor Thickness, mm	Histopathological Association	Skin Type	>50 Melanocytic Nevi	>1 Atypical Melanocytic Nevus
1/M/50	Upper back	II/<0.75	Atypical melanocytic nevus	III	–	+
2/F/50	Upper back	Total regression	NA	II	+	–
3/M/55	Calf	IV/1.60	Congenital compound nevus†	III	–	+
4/M/53	Upper back	II/<0.50	Congenital dermal nevus†	III	–	+
5/M/35	Upper back	IV/1.75	Atypical melanocytic nevus	II-III	+	–
6/M/56	Upper back	III/3.25	Congenital dermal nevus†	II	–	–
7/M/35	Thigh	III/1.25	Atypical melanocytic nevus	III	–	+
8/M/48	Upper back	III/1.60	Atypical melanocytic nevus	II	–	–

Abbreviations: MM, malignant melanoma; NA, data not available; +, present; –, absent.
*All patients had numerous solar lentigines.
†The associated congenital nevi were only evident on histopathological examination. Clinically, the overall diameters of the lesions ranged from 9 to 14 mm.

To study this in more detail, the researchers recruited 210 marathon runners (166 male, 44 female), all white. They also formed a control group of 210 white non-runners, matched by age and sex to the marathon running group. Each participant was given a questionnaire to identify risk factors (skin type etc.), then a full dermatological examination and skin cancer screening.

In the marathon group, 24 had potentially cancerous moles or lesions that were referred for further treatment; in the control group, only 14 did.

Is this a statistically significant difference?

Detailed results are in Table 2 of the paper.

Table 2. Distribution of Risk Factors for Malignant Melanoma in MG Compared With the Age- and Sex-Matched CG in This Study*

Risk Factor	Total MG (N = 210)	Total CG (N = 210)	Training Intensity of the MG, km/wk		
			<40 (n = 78)	40-70 (n = 101)	>70 (n = 31)
Anamnestic					
Personal or family history of skin cancer	**3** (1.4)	6 (2.9)	0	2 (2.0)	1 (3.2)
Changes in skin lesions	**28** (13.3)	36 (17.1)	10 (12.8)	16 (15.8)	2 (6.5)
Phenotypic					
Blond or red hair	**58** (27.6)	50 (23.8)	23 (29.5)	26 (25.7)	9 (29.0)
Blue, gray, or green eyes	**117** (55.7)†	141 (67.1)	41 (52.6)‡	56 (55.4)	20 (64.5)
Numerous ephelides	**53** (25.2)	58 (27.6)	25 (32.1)	23 (22.8)	5 (16.1)
High sun sensitivity (skin type I or II)	**114** (54.3)§	156 (74.3)	42 (53.8)§	56 (55.4)§	16 (51.6)‖
Sun related					
>10 Sunburns	**82** (39.0)	82 (39.0)	28 (35.9)	42 (41.6)	12 (38.7)
At least 1 sunburn with blisters	**46** (21.9)	59 (28.1)	20 (25.6)	15 (14.9)‖	11 (35.5)
Clinical					
>50 Common melanocytic nevi	**29** (13.8)‡	47 (22.4)	9 (11.5)	18 (17.8)	2 (6.5)
>1 Atypical melanocytic nevus	**99** (47.1)§	66 (31.4)	32 (41.0)	51 (50.5)§	16 (51.6)**¶**
Numerous solar lentigines	**64** (30.5)‖	42 (20.0)	23 (29.5)	28 (27.7)	13 (41.9)‖
Referral for excision	**24** (11.4)	14 (6.7)	5 (6.4)	13 (12.9)	6 (19.4)**¶**

Abbreviations: CG, control group; MG, marathon group.
*Data are given as number (percentages) of subjects.
†*P*=.02 vs CG.
‡*P*=.03 vs CG.
§*P*=.001 vs CG.
‖*P*=.01 vs CG.
¶*P*=.04 vs CG.

"Significant" results in the table are marked by P =.04, .03, .02, .01 or .001. It is worth pointing out that although a number of results in that table are flagged as statistically significant, the result about 24 v. 14 referrals for treatment is *not* flagged as statistically significant. However, among the "high training" group (more than 70km. per week), 6 out of 31 runners had to be referred and that *is* statistically significant, according to the table.

My conclusion was that both the newspaper report and the title of the article itself exaggerated what the study had actually proved. The study made a number of comparisons between the treatment and control groups, but most of them were not statistically significant, and some of them showed that it was the control group that was at greater risk.

However, this is also a small study. Despite the increase in skin cancer in recent years, it's still a relatively rare disease — well under 1% of the total population. In 420 subjects in the study, the researchers may well not have seen enough cases to make a meaningful comparison.

In this chapter, we discuss some general principles related to statistical significance and P-values that often come out in this sort of study.

Steps to performing a significance test

The melanoma example is actually a little different from the examples discussed in this chapter because it is about the *comparison* of two proportions — whether the proportion of a certain skin problem among marathon runners is higher than among the control group (in a situation where both proportions are unknown in the population at large). This is actually the subject of Chapter 10. For the purposes of Chapter 9, let's pretend that the control group was actually much larger, and that the 14/210 in the control group who had to be referred for treatment was actually representative of the whole population — 6.7% or $p = 0.067$. The question that then arises is whether either the proportion that had to be refereed for treatment among the marathon runner sample (24/210 or 11.43%), or the proportion among the "heavy trainers" (6/31 or 19.35%), are statistically different from the general population.

In the text they use an example related to astrology — an astrologer is given three possible personality profiles corresponding to a particular individual, and he/she has to guess which one is correct based on the individual's birth date. If there is no astrological effect, the proportion of correct guesses will be $p = \frac{1}{3}$. If there is an effect, presumably the astrologer will guess correctly greater than one-third of the time. As with the skin cancer example, the question is whether the observed proportion of a particular outcome in an experiment is significantly different from the proportion we would expect to see by chance if there was no effect.

Five steps to performing a significance test:

(1) Specify the assumptions. For example, many (if not most) studies require randomization.

(2) Define the hypotheses of interest. Typically, this kind of problem is formulated as a choice between the *null hypothesis* and the *alternative hypothesis*. The null hypothesis means there is no effect, e.g. the proportion of skin problems among marathon runners is no different from that of the general population, or the astrologer guesses correctly only one-third of the time. The alternative hypothesis is when the null hypothesis is not correct. The null hypothesis and alternative hypothesis are often written H_0 and H_a. So in the skin cancer example we may say that if p is the proportion of marathon runners referred for treatment,

$$\mathrm{H}_0 : p = 0.067, \qquad \mathrm{H}_\mathrm{a} : p > 0.067.$$

(3) Test statistic: Calculate some summary of the data that may be used to discriminate between the null and alternative hypotheses.

(4) P-value: Calculate the probability that a result, equal to or more extreme than the one actually observed, would occur if H_0 was correct. This is called the P-value (not to be confused with small p which represented the proportion we were trying to test).

(5) Report the conclusions. If the P-value is sufficiently small, we conclude that the null hypothesis is very unlikely to be correct and therefore conclude that the alternative hypothesis is correct.

Consider the skin cancer example applied to the "heavy training" group (6 out of 31 runners referred for treatment). In that case, these steps work out as follows:

(1) Assume that X, the number of marathon runners referred for treatment, has a binomial distribution with $n = 31$ and unknown p.

(2) The natural null and alternative hypotheses are $\mathsf{H}_0 : p = 0.067$ and $\mathsf{H}_\mathsf{a} : p > 0.067$.

(3) The test statistic is the sample proportion $\widehat{p} = \frac{6}{31} = 0.1935$.

If H_0 is correct, the standard error of $\widehat{p}$ is $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.067 \times 0.933}{31}} = .0449$. Therefore, the z statistic is

$$z = \frac{0.1935 - 0.067}{0.0449} = 2.82.$$

(4) Referring to the normal distribution table, the probability that a standard normal random variable is greater than 2.82 is .0024. Therefore, the P-value is .0024 in this case.

(5) The value .0024 is rather small — well under a 1% probability that this result could have occurred by chance. Therefore, we'd be justified in concluding that p is not 0.067 — in other words, the marathon runners really did have a higher incidence of this particular skin problem.

One comment here is that we used the normal approximation to the binomial distribution when one of the conditions for that to be value, that $np \geq 15$, is violated. In fact, if $n = 31$ and $p = .067$, then $np = 2.08$ so this is definitely wrong. In the present case, we could use the exact binomial distribution — if X has a binomial distribution with $n = 31$ and $p = 0.067$, then $\Pr\{X \geq 6\} = 0.0156$. This is a bit bigger than .0024, but the conclusion (that it's too small a probability to be attributed to chance) is still valid. In any case, most of the examples we will see do follow the normal distribution so we won't worry about that distinction.

General procedure for testing a proportion

Suppose the data are a sample proportion $\widehat{p}$ from the sample of size n where the true population proportion is an unknown quantity p. In the problems described in this chapter, the null hypothesis is always of the form

$$\mathsf{H}_0 : \ p = p_0$$

where p_0 is some given proportion. In the melanoma example, p_0 is 0.067, representing the proportion of people needing treatment in a control population (which in the actual example was only 210 people but, for the purpose of this discussion, we are assuming to represent the entire population). In the astrology example, p_0 is $\frac{1}{3}$, because if the astrologer had no special powers this would be her probability of guessing correctly in a single trial.

Defining the alternative hypothesis

The alternative hypothesis is almost always one of

$$H_a: \ p > p_0, \quad \text{or} \tag{1}$$

$$H_a: \ p < p_0, \quad \text{or} \tag{2}$$

$$H_a: \ p \neq p_0. \tag{3}$$

The choice among these depends on which alternative is more interesting in the context of the test.

In the astrology example, the clearest way to prove that the astrologer had real powers would be if the proportion of correct guesses was better than could be achieved by chance — in other words, if $p_0 > \frac{1}{3}$. The alternative $p_0 < \frac{1}{3}$ would not seem to make much sense. Therefore, in this case (1) would seem to be the natural choice.

Skin cancer example: it seems logical to argue the runners would be at an increased risk of skin cancer because they spend more time out of doors. Therefore the natural alternative hypothesis would seem to be (1) again.

However, it's not actually so clear-cut, because we have seen there are some risk factors for which the non-runners are at higher risk than the runners. There are various plausible explanations for this. So in fact, the researchers always took (3) as the alternative (for all the tests they did).

Test statistic

After defining the null and hypothesis, we calculate the *test statistic*, typically

$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}. \tag{4}$$

Rationale: if H_0 is true, then $\widehat{p}_0$ has mean $\mu = p$ and standard deviation (or standard error) $\sigma = \sqrt{\frac{p_0(1-p_0)}{n}}$, so (4) is just the formula $z = \frac{x-\mu}{\sigma}$ (with $x = \widehat{p}$) in this context. If H_0 is true, then z has a standard normal distribution with mean 0 and standard deviation 1.

Computing the P-value

It matters which of the three alternatives (1)—(3) is being used. For (1), compute right-tail probability: $\Pr\{Z > z\}$ where Z is a standard normal random variable and z is the number computed in (4). In the skin cancer example, z was 2.82 and the corresponding right-tail probability was 0.0024.

For the alternative (2), this is usually of interest only if $z < 0$ but then the probability to be calculated is $\Pr\{Z < z\}$, the mirror image of (1).

However if the alternative is (3), in this case we do something different. We need to consider extreme events on both sizes, which means $Z > |z|$ or $Z < |z|$. In most cases this simply means doubling the one-sided P-value.

So for example, in the skin cancer example, the two-sided probability would be $\Pr\{Z > 2.82\} + \Pr\{Z < -2.82\}$. However, by the symmetry of the normal distribution, $\Pr\{Z > 2.82\}$ and $\Pr\{Z < -2.82\}$ are the same and equal to .0024, so $\Pr\{Z > 2.82\} + \Pr\{Z < -2.82\} = 2 \times .0024 = .0048$. This is the correct P-value in this instance.

Interpreting the P-value

The general principle is: the smaller the P-value, the less likely it is that the null hypothesis is true. However that begs the question of "how small is small?"

We don't normally consider a result "significant" unless P$<$ 0.05. There is no specific reason why the cut-off is this — we could equally have taken the rule as P$<$ 0.1 or P$<$ 0.03 or something else — but 0.05 has become generally accepted as the standard.

However, it's also true that P=0.05 is not particularly strong evidence against H_0 — if this were our universal rule, we would still end up rejecting H_0 5% of the time when H_0 is true. If we knew that P was much smaller than 0.05, that would strengthen our belief that H_a is really correct.

Many researchers do the following:

1. If P$>$ 0.05, simply report that the result is not significant and leave it at that.

2. If P$\leq$ 0.05, report the exact value of P so that the reader can judge for him/herself just how significant a result it is.

This is what the researchers in the skin cancer paper did. In their Table 2, all cases where P$>$ 0.05 were left unlabelled as not statistically significant. However, when P$<$ 0.05, they marked it according to P=0.04, 0.03, 0.02, 0.01 or 0.001, whichever was closest to the actual value of P.