Lecture 9: Statistical thinking

Matti Pirinen

13.9.2022

In these notes, we go through a couple of topics where it is important to incorporate statistics in our ways of thinking.

Let's start from two concepts from probability calculus, namely independence and conditional probability.

Independence

- **Intuition**: Random variables X and Y are independent (in a population) if knowing the value of one of them does not tell anything more about the value of the other than what we can already learn by the population distribution of the other.
- **Example 9.1**: Suppose that X and Y are two individuals from the Finnish population and we haven't yet measured their cholesterol values. Our best guess for each of them is the population mean that we assume known. Suppose we then measure X and get $X = 2.7$. If Y is completely unrelated and unconnected to X, then the fact that we know value of X doesn't change our estimate for Y that stays at the population mean. In this case, X and Y are independent. However, if Y is the twin brother of X, then we expect that there is some similarity (more accurately, positive correlation) between values of X and Y and hence we consider that Y is likely to be to the same direction from the population mean as X was. Now X and Y are not independent: knowing value for one of them also tells something about the other on top of the population distribution of the values.
- **Definition**: X and Y are independent if and only if $P(X, Y) = P(X)P(Y)$, that is, their joint distribution is the product of their marginal distributions.

If we have a sequence of independent events (such as a sequence of 4 tosses of a fair coin), then the probability of the whole sequence (such as head-head-tails-head) is the product of the individual probabilities of the events (such as $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$). This is a simple and powerful way to compute probabilities for a sequence of multiple events by using the probabilities of the individual events. But be careful not to to apply it also to cases where events are not independent! For example, if events are positively correlated, then this multiplication rule will often grossly underestimate the probability of the observed events and lead, for example, to much too low P-values under the null hypotheses and hence to false rejections of null hypotheses.

Case of Sally Clark. In 1999, a 35-year old Sally Clark was convicted of the murder of her two children. Her first child died in 1996 and the death was then considered being due to "Sudden Infant Death Syndrome (SIDS)". Once her 2nd child also died some years later, police started to investigate also the first death and eventually Sally Clark was accused of two murders. A prominent part of the prosecution was a statistical claim based on an estimated probability of SIDS being 1/8543 of live births. A medical expert then claimed that the probability that two cases of SIDS would happen naturally in the same family was (1/8543)x(1/8543)=1/73,000,000, which he considered such a small number that it was evidence for Clark's guilty. This calculation obviously assumed independence between two SIDS in the same family. This statement went uncorrected through the court, while it is obvious that there can be genetic and other factors that are associated with SIDS and those factors are shared between the children in the same family. Hence the SIDS status of children in the same family are NOT independent, but rather positively correlated and the true probability of two SIDS occurring naturally would then likely be much larger that 1 in 73 million. Additionally, no matter how small the probability of two natural SIDS in one family might be, in court, that value in isolation should not be interpreted as evidence for Clark's guilt. Rather, the possibly small likelihood of two SIDS would need to have been compared against the likelihood that Clark had indeed murdered her two children, which also is very small, unless there is some direct evidence for the murders. Eventually, Sally Clark was set free a few years after she was considered innocent after appeals. More of the story here.

Conditional probability

When events are not independent, we need to consider *conditional probabilities*. We denote by $P(Y \,|\, X)$ the probability that Y happens *once we know that X has happened*. This is "conditional probability of Y given X". Events X and Y are independent if and only if $P(Y \,|\, X) = P(Y)$, that is, if and only if probability of Y does not depend on the occurrence of X.

Example 9.2. Consider migraine in adult population. Assume 20% of women have migraine and 5% of men have migraine. If 50% of population are males (and 100%-50% = 50% females), what is the probability that an individual from the population has migraine?

Intuitive answer: Take 1000 individuals from the population. 500 of them are males and 500 are females. 20% of the 500 females have migraine, that is, 100 females. 5% of the 500 males have migraine, that is, 25 males. Altogether 100+25=125 individuals have migraine. Prevalence of migraine in population would be 125 per 1000, or 12.5%.

Probability calculus:

- $P(M) = 0.5$ probability of being male,
- $P(F) = 1 - P(M) = 0.5$ probability of being female,
- $P(D \,|\, F) = 0.2$ probability of disease *conditional on* being female,
- $P(D \,|\, M) = 0.05$ probability of disease *conditional on* being male,
- $P(D)$ probability that an individual has the disease.

Two important properties of probability calculus are present in the following formula:

$P(D) = P(D \,\&\, M) + P(D \,\&\, F) = P(D \,|\, M)P(M) + P(D \,|\, F)P(F) = 0.05 \cdot 0.5 + 0.20 \cdot 0.5 = 0.125.$

Namely, the law of total probability splits $P(D) = P(D \,\&\, M) + P(D \,\&\, F)$ into sum of two non-overlapping (males vs females) parts, and the joint probability of two events $P(D \,\&\, M) = P(D \,|\, M)P(M)$ can always be expressed through conditional probability of the 1st event conditioned on the 2nd event times the probability of the 2nd event.

Bayes rule

Let's consider two events X and Y. (They can be, e.g., X="has migraine" and Y="is male".) Let's write out their joint probability using conditional probability two ways around:

$$P(X \,\&\, Y) = P(X)P(Y \,|\, X) = P(Y)P(X \,|\, Y).$$

From this we can solve

$$P(Y \,|\, X) = \frac{P(Y)P(X \,|\, Y)}{P(X)} = P(Y)\frac{P(X \,|\, Y)}{P(X)}.$$

This formula is called **Bayes rule** (or **Bayes formula**), and it tells how the overall probability of event Y will change if we have learned that X has happened. That is, probability of Y, $P(Y)$ needs to be multiplied by a factor $\frac{P(X \,|\, Y)}{P(X)}$ in order to get probability of Y *conditional on X*.

With Bayes rule we can see work out that probability that a migraineur is male is $P(M\,|\,D) = P(M)\frac{P(D\,|\,M)}{P(D)} = 0.5\frac{0.05}{0.125} = 0.2$.

Example 9.3. Consider a diagnostic test of a disease D that has sensitivity of α and specificity of β. What is the probability that an individidual who in a population screening test positive (+) truly has the disease, when the prevalence of the disease is p?

- $P(+\,|\,D) = \alpha$ (sensitivity is α)
- $P(+\,|\,\text{no}D) = 1-\beta$ (specificity is β)
- $P(D) = p$ (prevalence is p)
- $P(+) = P(+\,|\,D)P(D) + P(+\,|\,\text{no}D)P(\text{no}D) = \alpha p + (1-\beta)(1-p)$

Bayes rule says that the **positive predictive value** (PPV) is

$$P(D\,|\,+) = P(D)\frac{P(+\,|\,D)}{P(+)} = \frac{\alpha p}{\alpha p + (1-\beta)(1-p)}.$$

Take a test that is 99% sensitive and 99% specific. If the disease has prevalence 1/1000, what is the probability that a positive test implies disease?

```
a = 0.99
b = 0.99
p = 0.001
a*p/(a*p+(1-b)*(1-p))
```

```
## [1] 0.09016393
```

How is it possible that only 9% of positives have the disease when test captures well (99%) both the true positives and the true negatives? Let's consider 10000 individuals. Out of them, 10 has D. Assume all those 10 test positive due to high sensitivity. 9990 does not have D. Out of them 1%, i.e., ~100 still test positive. Thus, out of all 110 positive tests, only 10 (~9%) were true positives and 100 (~91%) were false positives.

The key is that while the specificity sounds high (99%), it still leads to many false positives when large groups of (healthy) people are screened. And since disease is rare, population screening screens mainly healthy people.

This topic was in the news in recent years in relation to the population scale corona virus testing and how to interpret its results. HS says "Mitä harvemmalla on ollut koronaviruksen aiheuttama covid-19-sairaus, sitä enemmän vasta-aineita verestä seulova testi antaa virheellisiä positiivisia tuloksia."

Examples 9.4.

1. Plot positive predictive value $P(D\,|\,+)$ as a function of (a) prevalence from 0.001 to 0.5 for sensitivity 99% and specificity 99%. Plot another curve with specificity of 99.9%.

2. What is a way to increase PPV of a particular diagnostic test?

Interpreting significant findings

Is a statistically significant finding a true positive or a false positive? Note that P-value does **not** answer this question! P-value is probability of observing certain kind of data sets, not a probability that a hypothesis is true.

Denote by S event of getting significant result with significance threshold α and denote by N the event that the null hypothesis is true and by A the complementary event that the null hypothesis is false (and alternative hypothesis is true). We have that $P(A) = 1 - P(N)$. We are interested in $P(A|S)$, i.e., probability that there is a non-zero effect given that we observe a significant test result. Bayes rule gives

$$P(A|S) = \frac{P(A)P(S|A)}{P(S)} \text{ and}$$

$$P(N|S) = \frac{P(N)P(S|N)}{P(S)}.$$

By dividing the first equation by the second we have

$$\frac{P(A|S)}{P(N|S)} = \frac{P(A)P(S|A)}{P(N)P(S|N)}.$$

This says that the odds of there being a non-zero effect, after we have observed a significant P-value, are the prior odds of a non-zero effect $(P(A)/P(N))$ times the ratio of probabilities of getting a significant result under the alternative model vs. the null model. By definition, $P(S|N) = \alpha$, i.e., under the null we get significant results with probability α. The term $P(S|A)$ is **statistical power** of the study to observe a non-zero effect. Thus,

$$\frac{P(A|S)}{P(N|S)} = \text{prior-odds} \times \frac{\text{power}}{\text{significance threshold}}.$$

If we assume that we have a well-powered study to detect effects we are interested in, say power is above 80%, we can replace power by ≈ 1 and ignore it. We see that whether a significant result is more likely to be a true positive than a false positive depends on the ratio of prior-odds of true effect and significance threshold. If we want our inference procedure to produce significant results only for almost certain cases of true positives, we need to choose our significance threshold small enough that it can overcome a possible small prior odds of a true effect. (Note, however, that power will also drop when we decrease the significance threshold so we cannot ignore it forever.)

Example 9.5. Suppose that we are looking for common genetic variants that increase the odds of heart disease by at least 20% compared to the other allele at that position of genome. We think that there are not many such variants around, maybe only 10 or so among the 10^7 common variants. Thus we say that our prior probability that any one variant is this kind of a risk variant is $P(A) = 10/10^7 = 10^{-6}$. What should be our significance threshold if we want to be over 95% certain that an (average) significant finding is truly a real effect? (Here "average" because this consideration does not consider the exact properties of the variant but only average properties of those variants that reach significance threshold.)

We have that

$$\text{significance threshold } \alpha = \text{power} \times \frac{\text{prior-odds}}{\text{posterior odds}} \leq \frac{\text{prior-odds}}{\text{posterior odds}}.$$

We don't know power as it depends on the significance threshold α, but we can bound power from above by 1, and hence get an upper bound for α.

```r
p.A = 1e-6
prior.odds = p.A/(1-p.A)
pwr = 1 #upper bound for power --> upper bound for alpha
post.odds = 0.95/(1-0.95)
alpha = prior.odds*pwr/post.odds
paste(signif(alpha,3))
```

```
## [1] "5.26e-08"
```

This is close to the commonly used genome-wide significance threshold 5×10^{-8}.

If we instead use significance threshold 0.05, then we have that posterior odds of a non-zero effect of an average significant finding is

```
prior.odds*pwr/0.05
```

```
## [1] 2.000002e-05
```

or 1/50000. This shows why significance threshold of 0.05 would not make sense in genome-analyses where prior probability of a true effect is tiny: Nearly all of our “significant findings” would be false positives.

An important message is thus that to interpret a P-value with respect to the evidence that the effect is real, we must also consider how likely a real non-zero effect was, a priori, before we saw the data or the P-value. For effects that are quite plausible a priori, a higher P-value may suffice to give us the same evidence in favor of a real effect than what we have with much smaller P-value for an effect that was very improbable a priori.

Multiple testing

Suppose that you study the effect of vitamin X on symptoms of disease D and get a P-value of 0.05 for the difference between the treatment groups. You write a paper and publish it in a scientific journal. Can the readers of the journal then interpret your result in such a way that if there was no effect, then only in 1 out of 20 cases would they be reading such a report?

What if there were 9 other researchers elsewhere in the world who also conducted a similar study but did not get a significant result at $P = 0.05$. How would that affect the interpretation?

Our null hypothesis is that there is no effect of X on D. We, as a community, do 10 independent studies. What is the probability that at least one study reports $P \leq 0.05$ under the null hypothesis? Is it 0.05? No, it is

$$\text{Pr}(\text{at least one } P \leq 0.05 \,|\, \text{NULL}) = 1 - \text{Pr}(\text{all } P > 0.05 \,|\, \text{NULL}) = 1 - 0.95^{10} \approx 0.401.$$

Thus the type I error rate is 40.1% rather than 5%! So even though there is only 5% probability for any one of the studies to report a false positive, taken the studies together, there is over 40% probability that at least one of the studies does so.

This is a **multiple testing problem**, where many tests are completed, but only the significant ones are reported. In that case, a correctly calibrated interpretation of the P-value would require accounting for all the relevant studies. See also: https://xkcd.com/882/.

Similar issue comes up when a single researcher studies many possible predictors. It would then be misleading if the researcher tests, say, 100 variables for association with a disease, but only reports the ones that were “significant” without reporting that the other tests were also done but were “not significant”.

A standard way to adjust the significance thresholds for the number of tests is the Bonferroni correction that states that when we do n tests, we should apply a significance level α/n to each of them in order to keep our overall type I error rate at α level. Thus, if one does 10 tests and wants to have an overall type I error at 0.05, meaning that only in 5% of the sets of 10 tests would we report at least one significant finding under the null hypothesis that none of the effects tested was real, then one should consider results significant only if $P < 0.005 = 0.05/10$.

Proportion of true positives

Let’s recall the formula for the odds of a significant P-value indicating a true positive rather than a false positive:

$$\frac{P(A|S)}{P(N|S)} = \frac{P(A)P(S|A)}{P(N)P(S|N)} = \text{prior-odds} \times \frac{\text{power}}{\text{significance threshold}}.$$

Thus, for a fixed significance threshold and prior-odds of association, the probability of a significant result being true increases proportional to the power of the study. Hence, a larger proportion of the significant findings from a well powered study is likely to be true positives than from an underpowered study. Another way to think this is that all studies (independent of their power) have the same rate of labelling null effects as significant (and this rate is α, the significance level, the Type I error rate), but different rates of labelling true effects as significant (and this rate is the power of the study). Hence, by increasing power, a larger proportion of all significant results will be true positives. This is why well-powered studies are important.

Significance testing vs. effect sizes

P-values have had a large role in the statistical inference we have considered. However, in many real applications, the null hypothesis testing should not be the main goal of inference because we know already beforehand that the null hypothesis is not true. For example, if we want to know how best to invest money to very different types of treatments, it is not useful to test whether the two treatments have exactly the same effect on patients: We know that they won't; the important question is how large is their difference.

In many applications we should be primarily interested in the value of the effect size, rather than the probability whether the effect is zero, let alone on its P-value. Furthermore, since almost all effects are non-zero when we look carefully enough, it follows that if we simply increase our sample size, we will see a statistically significant difference virtually in every possible comparison we can think of, at least in fields that study complex phenomena that are affected by almost an infinite number of factors. In those cases, a statistically highly significant result may not be at all significant in real life. For example, suppose we can show that group A has a statistically significantly (P-value < 0.05) higher risk for hospitalization in next 5 years than group B. If this result has been achieved from a large sample of individuals (say millions), and the difference in risk is very small, say 1% increased risk, then this result is unlikely to be surprising: groups A and B are different is some identifiable ways (otherwise you could not tell who belong to A and who to B) and therefore we also expect *at least some small* differences between them in many other measures we can imagine. The real question is how large the difference is. If the difference between the groups turns out to be large enough, say 10%, then we may want to seek for an explanation for it. Also in clinical medicine, the main interest should be in the effect size: If an expensive/complicated treatment has a tiny difference in success rate compared to no treatment, it may not be worth it.

Regression towards the mean

What do you think about the following claims:

1. "The previous 4 tosses of a coin have ended up 'heads'. So I bet that the next toss will land 'tails' more likely than 'heads' because 'heads' and 'tails' should be 50%:50% in the long run.
2. "I got 6 correct in lottery this week so next week I'm likely to do worse."

The second claim describe a "regression towards the mean" phenomenon that states that if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement, and if it is extreme on its second measurement, it will tend to have been closer to the average on its first measurement.

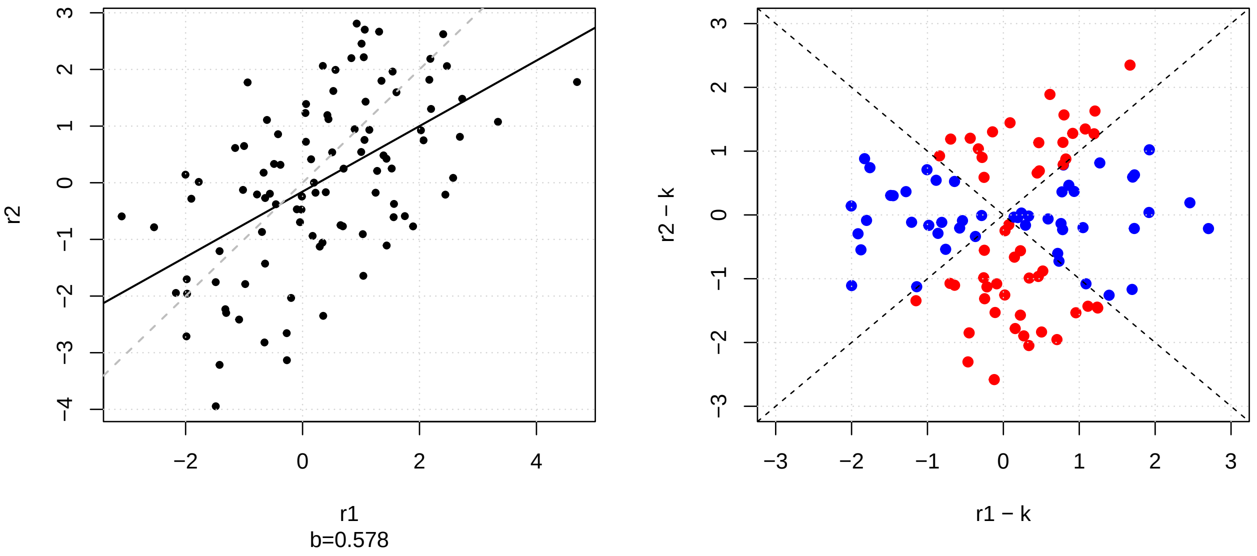
And the first claim is just wrong as the flips of a coin are independent events and heads and tails are equally probable outcomes. In lottery, it is overwhelmingly more probable to get less than 6 correct than to get at least 6 correct. Pay a careful attention to the difference between these two claims and why the first is false but the second is true.

Let's demonstrate regression towards the mean by simulating school children doing two exams on the same topic. Assume that the true level of knowledge k_i of each child i is Uniformly distributed on the interval (-2,2) in the population. Given the true knowledge, the exam result r_i has a distribution $\mathcal{N}(\text{mean}=k_i, \text{var}=1)$. The two exam results of child i are independent conditional on the knowledge k_i of the child.

```
n = 100
k = runif(n, -2, 2) # knowledge of each child
r1 = rnorm(n, k, 1) # result of exam 1
r2 = rnorm(n, k, 1) # results of exam 2
```

Let's plot the results against each other and regress `r2` on `r1`. Let's also show the absolute difference from the true value of k_i and color differently those children for whom `r1` is farther from the true value than `r2` compared to those where `r2` is farther.

```
par(mfrow = c(1,2))
lm.1 = lm(r2 ~ r1)
plot(r1, r2, pch = 20, sub = paste0("b=",signif(lm.1$coefficients[2],3)))
grid()
abline(lm.1, lwd = 1.5)
abline(0, 1, col = "gray", lty = 2, lwd = 1.5)
cols = rep("red", n) #will stay red if r1 is closer to true knowledge than r2
cols[abs(r1-k) > abs(r2-k)] = "blue" #blue if r2 is closer to true knowledge than r1
plot(r1 - k, r2 - k, col = cols, pch = 19, xlim = c(-3,3), ylim = c(-3,3))
grid()
abline(0, 1, lty = 2)
abline(0, -1, lty = 2)
```



We see that the regression line is below the diagonal (~0.6 here) which means that predicted value of `r2` based on `r1` is closer to the population mean of 0 than observed value of `r1`. Exactly the same happens if we regress `r1` on `r2`. This phenomenon of "regression towards the mean" is the original source for term "regression" in linear regression.

The second plot shows that most extreme deviations in both `r1` and `r2` are consistently less extreme in the other result. For example, look at the points that are < -2 or >2 on y-axis. They show a large deviation (above 2 units) from mean k in `r2`, and they are all colored red, denoting that their deviation from k is larger in `r2` than in `r1`.

Examples 9.6.

In light of the regression towards the mean, what do you say about the following claims.

1. “Extra rehersal after the first exam made the results of the worst performing students better whereas it worsen the results of the best performing students. Therefore, we should train only the worst performing students more but not the best performing students.”

2. “On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So please don’t tell us that reinforcement works and punishment does not, because the opposite is the case.”

3. “When the pain gets extreme, I take the medication and soon the pain decreases. When the pain has been moderate, I have also sometimes tried the medicine but haven’t noticed any difference. So the medicine does work but only for the pain that is extreme.”