Stat 200: Introduction to Statistical Inference **Winter 2023**

Lecture 1: Introduction and probability review

Lecturer: Dominik Rothenhäusler *January 9*

Note: *These lecture notes were written by Art Owen. If you like the material, he gets the credit! These notes have not been subjected to the usual scrutiny reserved for formal publications. They are meant as a memory aid for students who took stat 200 at Stanford University. They may be distributed outside this class only with the permission of Art Owen. Also, Stanford University holds the copyright.*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

1.1 Introduction

Welcome to stat 200, "Introduction to statistical inference". The prerequisite is stat 116 (probability). After that you can follow up with any of a great many MS level statistics courses. Our text is Rice (2007). There is a course syllabus on the web. Same for lots of other details.

Statistics began as the measure of the state. Rulers needed to know what they ruled. One of the oldest is a Babylonian tablet of agricultural productivity. There's an image in Stigler (2016). (That book is a fun read, but you are not expected to read it for this course.)

Over time statistics has oscillated between the problems of too much data (gotta summarize) and too little data (gotta squeeze out all the info we can). At present both take place.

Statistics is for "learning about the world from data". We might get data like the hypothetical example below. If those 280 people got randomly assigned to either the drug or the placebo then it looks like it was

	Cured	Not
drug	100	40
placebo	80	60

better to get the drug. However we probably are more about the millions of other people who might get that drug (or not) in the future. So, does the result here generalize further?

We have two problems. The data we saw are a sample, not the whole population we care about. Also, the data we saw could have been different. We handle both of those issues via probability.

Probability is a branch of mathematics built upon other branches of mathematics such as the theory of functions. Applied statistics involves the use of statistical methods to solve real problems in biology, engineering, education, economics, commerce and so on. Theoretical statistics sits in between probability and applied statistics studying ideas that apply broadly. I believe that statistics is philosophy that we can back up with mathematics. Sometimes we must use computers to do that mathematics. The key issues we face are those of deciding what to do under uncertainty, deciding what exactly we mean by uncertainty in a given problem, choosing models and judging to what extent we know what we think we know. Sometimes the

mathematical challenges are about lengthy calculations, and very rarely they are involve finding that one great trick. Almost always it is about what way of formulating the problem makes the most sense.

According to the Knudson hypothesis, cancer is the result of damage to DNA. In a one-hit model the damage arrives at a time $T \geqslant 0$ following a distribution with probability density function (PDF) $f_1(t) = \theta e^{-\theta t}$. This is an exponential distribution. We will see more of it later. It describes the time to some unpredictable 'bolt from the blue'. In a two-hit model the damage arrives at time $T = \max(T_1, T_2)$ where T_j are independent exponential random variables with parameters θ_j. Then

$$\begin{aligned}
\Pr(T \leqslant t) &= \Pr(T_1 \leqslant t, T_2 \leqslant t) && \text{because max} \\
&= \Pr(T_1 \leqslant t) \times \Pr(T_2 \leqslant t) && \text{independence} \\
&= (1 - e^{-\theta_1 t})(1 - e^{-\theta_2 t}) && \text{plug in.}
\end{aligned}$$

Then the two-hit PDF is

$$f_2(t) = \frac{\mathrm{d}}{\mathrm{d}t} Pr(T \leqslant t) = \cdots = \theta_1 e^{-\theta_1 t} + \theta_2 e^{-\theta_2 t} - (\theta_1 + \theta_2) e^{-(\theta_1 + \theta_2)t}$$

after some calculus. Statistical questions include

1. Do our data look more like one hit or two hit?
2. Even if we knew it was one hit, we would not usually know θ. So how would we estimate θ from data?
3. Surely there's more than one way to do that. Which one that we come up with is best?
4. Could there be a better estimate than the one we got, or is it the best possible?
5. How accurate is our estimate of θ?
6. How would we estimate (θ_1, θ_2) for two hit?
7. If we are planning this, how much data do we need to get?

1.2 Probability review

We began to review ideas from probability. These should already be familiar but a refresher is in order. Also not everybody had the same probability course. We will do statistics from Chapters 8 and on of Rice, so Chapters 1–6 of Rice are very tuned to that. (Chapter 7 is on sampling finite populations; we will skip that. Stat 204 covers it.)

We begin with a sample space Ω. This is a set of possible outcomes. The actual outcome is $\omega \in \Omega$. An event is a subset $A \subset \Omega$. We say that A occurs if and only if $\omega \in A$. At this point, think back to examples you learned about dice or coins being tossed. For a single coin $\Omega = \{H, T\}$ and $A = \{H\}$ is the event that the coin came up heads. Then probability is a function P on subsets[1] of Ω.

Probability is a measure of how likely an event to happen. What we mean by likely is context dependent. It can be strength of our opinion about event A. It can come from enumerating equally likely outcomes, counting how many are in A and dividing by the number in Ω, when Ω has a finite number of elements. The coin tossing, dice tossing and card sampling events you have studied are of this type. Or it could be the long run frequency with which A occurs in a conceptually infinite set of replications. We write it as $\Pr(A)$. Whatever it means, the following rules apply:

[1] Technical point that we will not use in this course: Sometimes Ω is complicated enough that only certain subsets can be assigned a probability.

1) $\Pr(\Omega) = 1$,
2) If $A \subseteq \Omega$ then $\Pr(A) \geqslant 0$, and
3) If A_i for $i = 1, 2, 3, \cdots, I$ satisfy $A_i \cap A_j = \varnothing$ whenever $i \neq j$, then $\Pr(\cup_{i=1}^{I} A_i) = \sum_{i=1}^{I} \Pr(A_I)$.

Rule 3 also holds for infinitely many sets A_i.

From these rules we get $\Pr(A) \leqslant 1$, $\Pr(A^c) = 1 - \Pr(A)$, $\Pr(\varnothing) = 0$ and $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$. Probability is a 'measure' just like length, mass, area, volume and so on (except they don't have to satisfy rule 1). This is why Venn diagrams work so well for them.

A critically important concept is conditional probability. If $\Pr(B) > 0$ then we say that the probability of event A happening given that event B has happened is

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

We just say the probability of A given B. If we make a new function of events called $q(\cdot)$ with $q(\cdot) = \Pr(\cdot \mid B)$, meaning that $q(A) = \Pr(A \mid B)$ for any A, then q is also a probability. It obeys the rules we have laid out above for Pr. We can even have $q(\cdot \mid C)$ for some other event C with $q(C) > 0$.

Events A and B are independent if $\Pr(A \cap B) = \Pr(A) \times \Pr(B)$. In applications, independence is a powerful assumption that greatly simplifies things. It is easy to show that independent events A and B satisfy $\Pr(A \mid B) = \Pr(A \mid B^c) = \Pr(A)$ though we need $\Pr(B) > 0$ for the first one to be well defined and $\Pr(B) < 1$ for the second one.

1.3 Random variables

A random variable X is defined as a function on Ω. Then $X(\omega)$ is a real value. The event $X \in A$ is the same as $\{\omega \mid X(\omega) \in A\}$. We usually don't write $X(\omega)$ because that gets cumbersome, but the rules of probability for random variables come from their representation as functions on Ω.

The cumulative distribution function of a random variable X is $F(x) = \Pr(X \leqslant x)$, actually $\Pr(\{\omega \in \Omega \mid X(\omega) \leqslant x)$ but that gets cumbersome. Here X is the random variable and x is one particular value it might take. So $F(x)$ is a function on the real line. It satisfies $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$ and $F(x) \geqslant F(x')$ whenever $x \geqslant x'$ (monotonicity).

Some random variables are discrete, taking only values $x \in \{x_1, x_2, \dots\}$ a set which may be finite or infinite. If X is discrete with $\Pr(X = x_i) = p(x_i) > 0$ then $F(x)$ has a jump of size $p(x_i)$ at x_i and is continuous from the right. It may resemble a staircase when graphed.

Other random variables are continuous. They have a probability density function (PDF) $f(x)$. For such an rv and $a < b$,

$$\Pr(a < X < b) = \int_a^b f(x)\,\mathrm{d}x.$$

An important example for us is the normal (or Gaussian) distribution with mean μ and variance $\sigma^2 > 0$. It has PDF

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

The standard normal distribution has $\mu = 0$ and $\sigma = 1$. A PDF must satisfy $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$. If we knew that we wanted a PDF proportional to $e^{-x^2/2}$, that is $f(x) = ce^{-x^2/2}$ for some constant $c > 0$, then we would have had to solve an integral to realize that the constant of proportionality is $c = 1/\sqrt{2\pi}$. Very often we do the reverse. Knowing the form of the normal PDF allows us to know that $\int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x = \sqrt{2\pi}$.

If you replaced the standard normal PDF by

$$\tilde{f}(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} & x \neq 11 \\ 42 & x = 11, \end{cases}$$

then the resulting distribution would be unchanged. For instance $F(x)$ would not change and neither would $\Pr(a < X < b)$ for any $a < b$. You could even change the PDF f at any finite or countable number of points without changing the distribution. We won't be doing that. Also for a continuous random variable

$$\Pr(a < X < b) = \Pr(a < X \leqslant b) = \Pr(a \leqslant X < b) = \Pr(a \leqslant X \leqslant b).$$

If X is a random variable and $Y = g(X)$ then Y is also a random variable. The distribution of Y can in principle be found from that of X, and in special cases the math simplifies enough that we can do it explicitly. For instance, let $X \sim U(0,1)$ and $Y = -\log(X)$. Then X has the uniform distribution on the interval $(0,1)$ with PDF

$$f_X(x) = \begin{cases} 0 & x < 0 \\ 1 & 0 \leqslant x \leqslant 1 \\ 0 & x > 1. \end{cases}$$

and CDF

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leqslant x \leqslant 1 \\ 1 & x > 1. \end{cases}$$

We want the PDF of Y which we call f_Y to keep it separate from f_X. First we find F_Y in slow motion, keeping an eye out for X vs x vs Y vs y. Because $0 < X < 1$ we have $0 < Y < \infty$. So for $y > 0$,

$$\begin{aligned} F_Y(y) &= \Pr(Y \leqslant y) \\ &= \Pr(-\log(X) \leqslant y) \\ &= \Pr(\log(X) \geqslant -y) \\ &= \Pr(X \geqslant e^{-y}) \\ &= 1 - \Pr(X < e^{-y}) \\ &= 1 - \Pr(X \leqslant e^{-y}) \\ &= 1 - F_X(e^{-y}) \\ &= 1 - e^{-y}. \end{aligned}$$

Next

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_y(y) = e^{-y}.$$

More precisely

$$f_Y(y) = \begin{cases} 0 & y \leqslant 0 \\ e^{-y} & y > 0. \end{cases}$$

We certainly don't want to have $f_Y(-3) = e^3$, so handling cases is important for a complete description of a PDF. We could have ended up with $f_Y(0) = e^{-0} = 1$ instead of $f_Y(0) = 0$. It would be the same distribution.

If g is an increasing and invertible function then

$$F_Y(y) = \Pr(Y \leqslant y) = \Pr(g(X) \leqslant y) = \Pr(X \leqslant g^{-1}(y)) = F_X(g^{-1}(y)).$$

We also reviewed expectation of random variables but that will be folded into the next scribing.

1.4 Coda

Here's a fun quote from Larry Wasserman of Carnegie-Mellon.

> "Students who analyze data, or who aspire to develop new methods for analyzing data, should be well grounded in basic probability and mathematical statistics. Using fancy tools like neural nets, boosting, and support vector machines without understanding basic statistics is like doing brain surgery before knowing how to use a band-aid." –Larry Wasserman

At the time he wrote, support vector machines were a much more prominent way to do statistical machine learning than they are now. There is a lot to be said for using the most powerful and sophicated methods around. There is also a lot to be said for understanding exactly what you're doing top to bottom, knowing the use cases, assumptions, gotchas and special cases.

References

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis.* Brooks/Cole, Belmont, CA.

Stigler, S. M. (2016). *The seven pillars of statistical wisdom.* Harvard University Press.