Chapter 6: Wrapping Up

Probably the two major ideas of this course are:

- likelihood and estimation;
- hypothesis testing.

Most of the techniques that we have studied along the way are to help us with these two goals: expectation, variance, distributions, change of variable, and the Central Limit Theorem.

Let's see how these different ideas all come together.

6.1 Estimators — the good, the bad, and the estimator PDF

We have seen that an *estimator* is a capital letter replacing a small letter. What's the point of that?

Example: Let $X \sim \text{Binomial}(n, p)$ with known n and observed value $X = x$.

- The maximum likelihood *estimate* of p is $\widehat{p} = \frac{x}{n}$.
- The maximum likelihood *estimator* of p is $\widehat{p} = \frac{X}{n}$.

Example: Let $X \sim \text{Exponential}(\lambda)$ with observed value $X = x$.

- The maximum likelihood *estimate* of λ is $\widehat{\lambda} = \frac{1}{x}$.
- The maximum likelihood *estimator* of λ is $\widehat{\lambda} = \frac{1}{X}$.

Why are we interested in *estimators*?

The answer is that ***estimators are random variables.*** This means they have ***distributions, means,*** and ***variances*** that tell us how well we can trust our single observation, or ***estimate,*** from this distribution.

Good and bad estimators

Suppose that $X_1, X_2, \ldots, X_n$ are independent, and $X_i \sim$ Exponential(λ) for all i. λ is unknown, and we wish to estimate it.

In Chapter 4 we calculated the maximum likelihood estimator of λ:

$$\widehat{\lambda} = \frac{1}{\overline{X}} = \frac{n}{X_1 + X_2 + \ldots + X_n}.$$

Now $\widehat{\lambda}$ is a *random variable with a distribution.*

For a given value of n, we can calculate the p.d.f. of $\widehat{\lambda}$. How?

We know that $T = X_1 + \ldots + X_n \sim$ *Gamma*(n, λ) *when* $X_i \sim$ *i.i.d. Exponential*(λ).

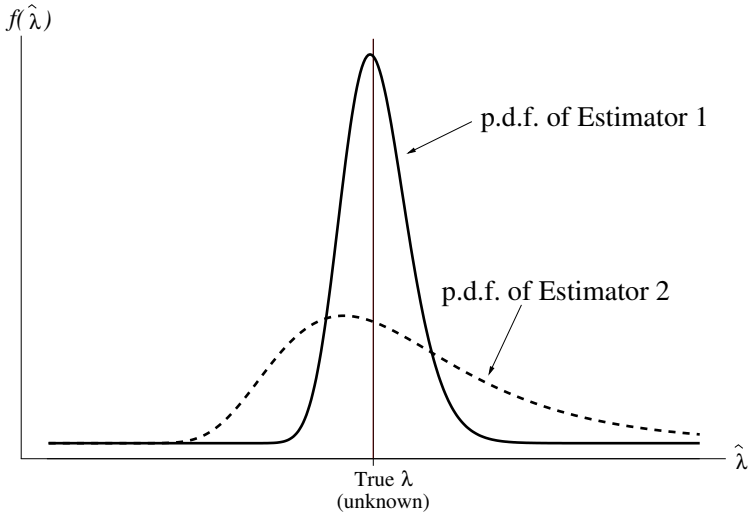
So we know *the p.d.f. of* T.

Now $\widehat{\lambda} = \frac{n}{T}$.

So we can find the p.d.f. of $\widehat{\lambda}$ using the *change of variable technique.*

Here are the p.d.f.s of $\widehat{\lambda}$ for *two different values of* n*:*

- Estimator 1: $n = 100$. *100 pieces of information about* λ.
- Estimator 2: $n = 10$. *10 pieces of information about* λ.



Clearly, the more information we have, the better. The p.d.f. for $n = 100$ is focused much more tightly about the true value λ (unknown) than the p.d.f. for $n = 10$.

It is important to recognise what we do and don't know in this situation:

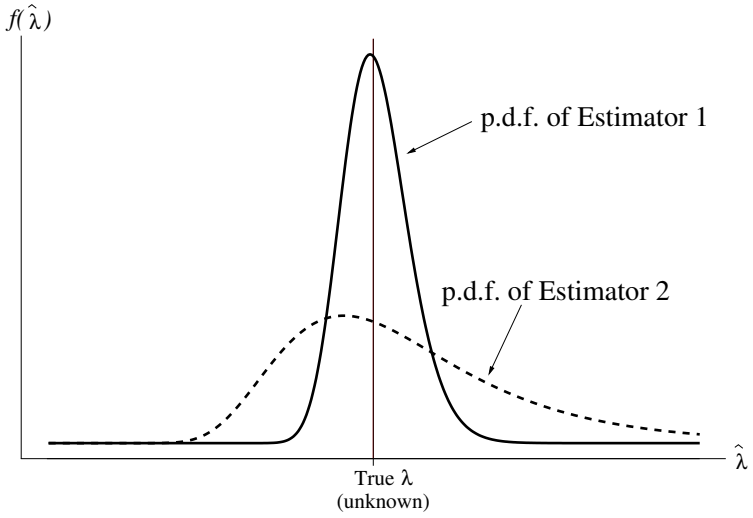
What we don't know:

- *the true λ;*
- *WHERE we are on the p.d.f. curve.*

What we do know:

- *the p.d.f. curve;*
- *we know we're SOMEWHERE on that curve.*

So we need an estimator such that *EVERYWHERE on the estimator's p.d.f. curve is good!*



This is why we are so concerned with *estimator variance.*

A ***good estimator*** has *low estimator variance:* everywhere on the estimator's p.d.f. curve is guaranteed to be good.

A ***poor estimator*** has *high estimator variance:* some places on the estimator's p.d.f. curve may be good, while others may be very bad. Because we *don't know where we are on the curve,* we can't trust *any* estimate from this poor estimator.

The estimator variance tells us how much the estimator can be trusted.

Note: We were lucky in this example to happen to know that $T = X_1 + \ldots + X_n \sim \text{Gamma}(n, \lambda)$ when $X_i \sim$ i.i.d. Exponential(λ), so we could find the p.d.f. of our estimator $\widehat{\lambda} = n/T$. We won't usually be so lucky: so what should we do? *Use the Central Limit Theorem!*

Example: calculating the maximum likelihood estimator

The following question is in the same style as the exam questions.

Let X be a continuous random variable with probability density function

$$f_X(x) = \begin{cases} \dfrac{2(s-x)}{s^2} & \text{for } 0 < x < s\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Here, s is a parameter to be estimated, where s is the maximum value of X and $s > 0$.

(a) Show that $\mathbb{E}(X) = \dfrac{s}{3}$.

Use $\mathbb{E}(X) = \displaystyle\int_0^s x f_X(x)\,dx = \frac{2}{s^2}\int_0^s (sx - x^2)\,dx.$

(b) Show that $\mathbb{E}(X^2) = \dfrac{s^2}{6}$.

Use $\mathbb{E}(X^2) = \displaystyle\int_0^s x^2 f_X(x)\,dx = \frac{2}{s^2}\int_0^s (sx^2 - x^3)\,dx.$

(c) Find $\mathrm{Var}(X)$.

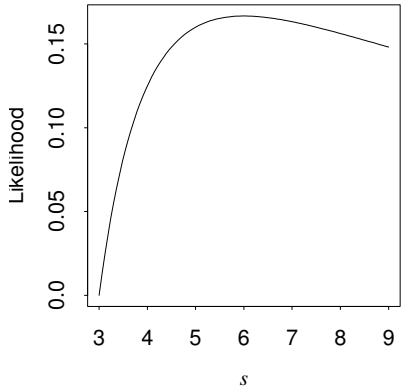
Use $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$. *Answer:* $\mathrm{Var}(X) = \frac{s^2}{18}$.

(d) Suppose that we make a single observation $X = x$. Write down the likelihood function, $L(s\,;\,x)$, and state the range of values of s for which your answer is valid.

$L(s\,;x) = \dfrac{2(s-x)}{s^2}$ *for* $x < s < \infty$.

(e) The likelihood graph for a particular value of x is shown here.

Show that the maximum likelihood estimator of s is $\widehat{s} = 2X$. You should refer to the graph in your answer.



$$\begin{aligned} L(s\,;\,x) &= 2s^{-2}(s-x) \\ \textit{So} \qquad \frac{dL}{ds} &= 2\left\{-2s^{-3}(s-x)+s^{-2}\right\} \\ &= 2s^{-3}(-2(s-x)+s) \\ &= \frac{2}{s^3}(2x-s). \end{aligned}$$

At the MLE,

$$\frac{dL}{ds}=0 \quad \Rightarrow \quad s=\infty \quad \text{or} \quad s=2x.$$

From the graph, we can see that $s=\infty$ is not the maximum. So $s=2x$.

Thus the maximum likelihood estimator is

$$\widehat{s}=2X.$$

(f) Find the estimator variance, $\mathrm{Var}(\widehat{s})$, in terms of s. Hence find the estimated variance, $\widehat{\mathrm{Var}}(\widehat{s})$, in terms of $\widehat{s}$.

$$\begin{aligned} \mathit{Var}(\widehat{s}) &= \mathit{Var}(2X) \\ &= 2^2\,\mathit{Var}(X) \\ &= 4\times\frac{s^2}{18} \qquad \textit{by (c)} \\ \mathit{Var}(\widehat{s}) &= \frac{2s^2}{9}. \end{aligned}$$

$$\textit{So also:} \qquad \widehat{\mathit{Var}}(\widehat{s}) = \frac{2\widehat{s}^2}{9}.$$

(g) Suppose we make the single observation $X = 3$. Find the maximum likelihood estimate of s, and its estimated variance and standard error.

$$\widehat{s} = 2X = 2 \times 3 = 6.$$

$$\widehat{Var}(\widehat{s}) = \frac{2\widehat{s}^2}{9} = \frac{2 \times 6^2}{9} = 8$$

$$se(\widehat{s}) = \sqrt{\widehat{Var}(\widehat{s})} = \sqrt{8} = 2.82.$$

This means $\widehat{s}$ is a POOR estimator: the twice standard-error interval would be $6 - 2 \times 2.82$ to $6 + 2 \times 2.82$: that is, 0.36 to 11.64 !

Taking the twice standard error interval strictly applies only to the Normal distribution, but it is a useful rule of thumb to see how 'good' the estimator is.

(h) Write a sentence in plain English to explain what the maximum likelihood estimate from part (g) represents.

The value $\widehat{s} = 6$ is the value of s under which the observation $X = 3$ is more likely than it is at any other value of s.

6.2 Hypothesis tests: in search of a distribution

When we do a hypothesis test, we need a ***test statistic:*** some random variable with a ***distribution*** that we can specify exactly under H_0 and that differs under H_1.

It is ***finding the distribution*** that is the difficult part.

- **Weird coin:** is my coin fair? Let X be the number of heads out of 10 tosses. $X \sim \text{Binomial}(10, p)$. We have an easy distribution and can do a hypothesis test.
- **Too many daughters?** Do divers have more daughters than sons? Let X be the number of daughters out of 190 diver children. $X \sim \text{Binomial}(190, p)$. Easy.

- **Too long between volcanoes?** Let X be the length of time between volcanic eruptions. If we assume volcanoes occur as a Poisson process, then $X \sim$ Exponential(λ). We have a simple distribution and test statistic (X): we can test the observed length of time between eruptions and see if it this is a believable observation under a hypothesized value of λ.

More advanced tests

Most things in life are not as easy as the three examples above.

Here are some observations. Do they come from a distribution (any distribution) with mean 0?

```
3.96  2.32 -1.81 -0.14  3.22  1.07 -0.52  0.40  0.51  1.48
1.37 -0.17  1.85  0.61 -0.58  1.54 -1.42 -0.85  1.66  1.54
```

Answer: yes, they are Normal(0, 4), but how can we tell?

What about these?

```
 3.3 -30.0  -7.8   3.4  -1.3  12.6  -9.6   1.4  -6.4 -11.8
-8.1   8.1  -9.0   8.1 -13.7  -5.0  -6.6  -5.6   2.5   9.0
```

Again, yes they do (Normal(0, 100) this time), but how can we tell? The unknown variance (4 versus 100) interferes, so that the second sample does not cluster about its mean of 0 at all.

What test statistic should we use?

If we don't know that our data are Normal, and we don't know their underlying variance, ***what*** can we use as our X to test whether $\mu = 0$?

Answer: a clever person called W. S. Gossett (1876-1937) worked out an answer. He called himself only 'Student', possibly because he (or his employers) wanted it to be kept secret that he was doing his statistical research as part of his employment at Guinness Brewery. The test that 'Student' developed is the familiar Student's t-test. It was originally developed to help Guinness decide how large a sample of people should be used in its beer tastings!

Student used the following test statistic for the unknown mean, μ:

$$T = \frac{\overline{X} - \mu}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n(n-1)}}}$$

Under $H_0 : \mu = 0$, the distribution of T is known: T has p.d.f.

$$f_T(t) = \left(\frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)} \right) \left(1 + \frac{t^2}{n-1} \right)^{-n/2} \qquad \text{for } -\infty < t < \infty.$$

T is the Student's t-distribution, derived as the ratio of a Normal random variable and an independent Chi-Squared random variable. If $\mu \neq 0$, observations of T will tend to lie out in the tails of this distribution.

The Student's t-test is exact when the distribution of the original data $X_1, \ldots, X_n$ is Normal. For other distributions, it is still approximately valid in large samples, by the Central Limit Theorem.

It looks difficult

It is! Most of the statistical tests in common use have deep (and sometimes quite impenetrable) theory behind them. As you can probably guess, Student did not derive the distribution above without a great deal of hard work. The result, however, is astonishing. With the help of our best friend the Central Limit Theorem, Student's T-statistic gives us a test for $\mu = 0$ (or any other value) that can be used with any large enough sample.

The Chi-squared test for testing proportions in a contingency table also has a deep theory, but once researchers had derived the ***distribution of a suitable test statistic,*** the rest was easy. In the Chi-squared goodness-of-fit test, the Pearson's chi-square test statistic is shown to have a Chi-squared distribution under H_0. It produces larger values under H_1.

One interesting point to note is the pivotal role of the Central Limit Theorem in all of this. The Central Limit Theorem produces approximate Normal distributions. Normal random variables squared produce Chi-squared random variables. Normals divided by Chi-squareds produce t-distributed random variables. A ratio of two Chi-squared distributions produces an F-distributed random variable. All these things are not coincidental: ***the Central Limit Theorem rules!***