# Stats 135, Fall 2019
Lecture 4, Friday, 9/6/2019

Last time, we showed the highlighted part of the formula table. To wrap up, Chapter 7 was about the population mean, estimating $\mu, p, \sigma^2$ and looking at the SE of them. Now in Chapter 8, we're going to generalize this. We have some data that fits a given distribution, and we need an estimator that should have some nice properties.
Further, we motivated why we estimate parameters of a probability model (as in the hospital Poisson example).

## 1 §8.4 Method of Moment (MOM) estimators

An estimator should converge to the true value (this is called **consistency**). There are different notions and definitions of convergence of a random variable (we will focus on the definition for Probability).
We'll first review the Gamma ($\Gamma$) distribution in the $\alpha$-particle example. Suppose that $X \sim \text{Gamma(r}, \lambda)$, where $X$ is the time to the $r$th arrival of a Poisson process. Here, $r$ is the $r$th particle, and $\lambda$ is the rate of arrival of $\alpha$-particles in the Poisson process. This has a density that we should know:

$$f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x},$$

and recall:

$$\Gamma(r) = (r-1)!,\ \ r \in \mathbb{Z}^+$$

and

$$\mathbb{E}(x) = \frac{r}{\lambda}$$
$$\text{Var}(x) = \frac{r}{\lambda^2}.$$

Now for Method of Moment estimators (MOM), if we want to estimate $l$ parameters $(\theta_1, \ldots, \theta_l)$ of a probability distribution $f(x \mid \theta_1, \ldots, \theta_l)$ from iid sample $x_1, \cdots, x_n$ from this distributions, there are 3 steps:

### 1.1 (Step 1)

We compute the first $l$ moments (where moments are the $k$th expectation):

$$\mu_k = \mathbb{E}(x^k), \quad k = 1, \ldots, l.$$

Now the RHS is given by the integral:

$$\mu_k = \int_{-\infty}^{\infty} x^k f(x \mid \theta_1, \ldots, \theta_l)\ dx.$$

This does depend on what these $\theta_i$ are. For $i$ in `1 : k`, we'll say these $\mu_i$ are functions $g_i$ on the arguments $\theta_j$. That is, we have the family of equations:

$$\begin{aligned}
\mu_1 &= g_1(\theta_1, \ldots, \theta_l) \\
\mu_2 &= g_2(\theta_1, \ldots, \theta_l) \\
&\vdots \\
\mu_l &= g_l(\theta_1, \ldots, \theta_l).
\end{aligned}$$

**Example:** Let $X \sim \text{Poisson}(\lambda)$, with $l = 1$ and let $X$ be the number of arrivals in 10 second intervals. The average rate of arrivals is just $\lambda$:

$$\mu_1 = \mathbb{E}(x) = \lambda.$$

**Example:** Suppose $X \sim \text{Gamma}(\text{r}, \lambda)$ now with $l = 2$. Then,

$$\begin{aligned}
\mu_1 &= \mathbb{E}(x) = \frac{r}{\lambda} \\
\mu_2 &= \mathbb{E}(x^2) = \underbrace{\text{Var}(x)}_{r/\lambda^2} + \underbrace{\mathbb{E}(x)^2}_{(r/\lambda)^2} = \frac{r + r^2}{\lambda^2}.
\end{aligned}$$

## 1.2 (Step 2)

Now we use algebra to invert the above system of equations (require $h$ to be a continuous function of $\mu_1, \ldots, \mu_l$):

$$\begin{aligned}
\theta_1 &= h_1(\mu_1, \ldots, \mu_l) \\
\theta_2 &= h_2(\mu_1, \ldots, \mu_l) \\
&\vdots \\
\theta_l &= h_l(\mu_1, \ldots, \mu_l)
\end{aligned}$$

**Example:** In the Poisson case, then we simply have:

$$\mu_1 = \lambda \implies \lambda = \mu_1.$$

We wrote $\mu_1$ in terms of the parameter, and in step 2 we wrote the parameter in terms of the moment. Done!

**Example:** In the Gamma case, we have:

$$\begin{aligned}
\mu_1 &= \frac{r}{\lambda} \\
\mu_2 &= \frac{r}{\lambda^2} + \frac{r^2}{\lambda^2} = \frac{\mu_1}{\lambda} + \mu_1^2 \\
\implies \frac{\mu}{\lambda} &= \mu_2 - \mu_1^2,
\end{aligned}$$

so this gives:

$$\begin{aligned}
\lambda &= \frac{\mu_1}{\mu_2 - \mu_1^2} \\
r = \lambda\mu_1 &= \frac{\mu_1^2}{\mu_2 - \mu_1^2}
\end{aligned}$$

## 1.3 (Step 3)

Now we insert into $(*)$ the estimator for the moments $\mu_1, \ldots, \mu_l$. We call these **sample moments**.

The first moment is the mean, so the first sample moment is the sample mean:

$$\begin{aligned} \hat{\mu}_1 &= \frac{1}{n}\sum_{i=1}^{n} x_1 \\ \hat{\mu}_2 &= \frac{1}{n}\sum_{i=1}^{n} x_1^2 \\ \mathbb{E}(x^2) \qquad &\vdots \\ \hat{\mu}_l &= \frac{1}{n}\sum_{i=1}^{n} x_i^l. \end{aligned}$$

We will show in homework that these are unbiased estimators of $\mu_1, \dots, \mu_l$. Now we have:

$$\begin{aligned} \hat{\theta}_1 &= h_1(\hat{\mu}_1, \dots, \hat{\mu}_l) \\ \hat{\theta}_2 &= h_2(\hat{\mu}_1, \dots, \hat{\mu}_l) \\ &\vdots \\ \hat{\theta}_l &= h_l(\hat{\mu}_1, \dots, \hat{\mu}_l), \end{aligned}$$

where we essentially just replace the non-hats with hats. We call these the **MOM estimators** for $\theta_1, \dots, \theta_l$. In practice, these are usually not the best estimators, but they are simple (just algebraic) so we talk about it now.
For example, we have:
Poisson:

$$\lambda = \mu_1 \implies \hat{\lambda} = \hat{\mu}_1 = \hat{x}$$

and for Gamma:

$$\begin{aligned} \hat{\lambda} &= \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2} \\ \hat{r} &= \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2} \end{aligned}$$

## 2 Showing MOM estimators are Consistent

This is the most fundamental property that we want, which is to say that

$$\hat{\theta}_{MOM} \xrightarrow{p} \theta,$$

where we write $p$ to mean convergence in the probability sense.
We first take a short digression to prove the **Weak Law of Large Numbers**. Our book doesn't go into this, so Lucas wants us to have this little missing piece.
We want to have Markov's inequality:

**Theorem 2.1.** (Markov's Inequality) :
For $x \geq 0$, $c > 0$, then the tail probability has the bound:

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(X)}{c}.$$

We won't prove this in lecture (Adam diverts the proof to Pitman).

**Example:** Let $x_1, \ldots, x_n$ be iid with mean $\mu$ and variance $\sigma^2$. Then the sample mean is:

$$\overline{x}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

and note that we know this is unbiased and we know the variance:

$$\mathbb{E}(\overline{x}_{(n)}) = \mu, \qquad \mathrm{Var}(\overline{x}_{(n)}) = \frac{\sigma^2}{n}.$$

Let $\epsilon > 0$. Use Markov's inequality to give an upper bound for

$$\mathbb{P}( \underbrace{|\overline{x}_{(n)} - \mu|}_{\text{nonrandom var}} \geq \underbrace{\epsilon}_{c>0})$$

So we have:

$$\begin{aligned}
\mathbb{P}(|\overline{x}_{(n)} - \mu| \geq \epsilon) &= \mathbb{P}[(\overline{x}_n - \mu)^2 \geq \epsilon^2] \\
&\leq \frac{\mathbb{E}\left[(\overline{x}_{(n)} - \mu)^2\right]}{\epsilon^2} \\
&= \frac{\mathrm{Var}(\overline{x}_{(n)})}{\epsilon^2} \\
&= \frac{\sigma^2}{n\epsilon^2},
\end{aligned}$$

where in the first line we square both sides because both are positive, and we notice we have $n$ in the denominator, so taking a distribution of $\overline{x}_{(n)} - \mu$ is centered at 0 and the tail area gets smaller as $n \to \infty$. More precisely, the area is bound by:

$$\frac{\sigma^2}{n\epsilon^2} \to 0, \text{ as } n \to \infty.$$

**Definition: Convergence in Probability -**

In friendly words, this is the idea that the probability of an unusual outcome gets smaller and smaller as $n$ grows.
To say that one random variable converges in probability to another, as $n$ grows larger, it will be very unlikely that their difference will be any greater, than say $\epsilon$.
For example, take a sequence $X_1, X_2, \ldots$ of random variables. We say this sequence converges in probability to a random variable $X$ if $\forall_{\epsilon>0}$, the probability:

$$\mathbb{P}(|x_n - x| > \epsilon) \to 0, \text{ as } n \to \infty.$$

**Example of Weak Law of Large Numbers:** The Weak Law of Large Numbers says that for $x_1, \ldots, x_n$ iid with mean $\mu$ and variance $\sigma^2$, we have:

$$\overline{x}_{(n)} \xrightarrow{p} \mu,$$

where $\mu$ is the constant random variable that takes the value $\mu$ with probability 1.

This follows directly from Markov's inequality.
We already derived:

$$\mathbb{P}(|\overline{x}_{(n)} - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2 \cdot n} \to 0, \text{ as } n \to \infty.$$

This proves the Weak Law of Large numbers. We can generalize this to higher moments. Take $x_1, \ldots, x_n$ iid with mean $\mu$ and variance $\sigma^2$. This says:

$$\hat{\mu}_k \xrightarrow{p} \mu_k,$$

where

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} x^k, \text{ and } \mu_k = \mathbb{E}(x^k).$$

Now we want to show consistency.

**Definition: Consistency -**

An estimator $\hat{\theta}$ of a parameter $\theta$ is **consistent** if $\hat{\theta} \xrightarrow{p} \theta$.

We argue that MOM is consistent. Adam Lucas notes that there is a highly-believable theorem:

**Theorem 2.2.** Suppose random variables $x_1, x_2, \ldots$ converge in probability to a random variable $x$ and $h$ is a **continuous** function. Then $h(x_1), h(x_2), \ldots$ converge in probability to $h(x)$.

Lucas states that to make our weekend complete, consider that if $h$ is continuous (as in the Method of Moments case), then the estimator is **consistent**. In other words, our MOM estimator:

$$\hat{\theta}_{MOM} = h(\hat{\mu}_1, \ldots, \hat{\mu}_l) \xrightarrow{p} \theta = h(\mu_1, \ldots, \mu_l),$$

which is very clear to see.

Lecture ends here.