STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 11

- If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables from $N(\mu, \sigma^2)$, then

$$\frac{(n-1)s^2}{\sigma^2} = \frac{n\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

- We have showed that the MLE for μ is $\hat{\mu} = \bar{X}$, which has a Normal$(\mu, \sigma^2/n)$ distribution.
- What is the (joint) sampling distribution of $(\hat{\mu}, \hat{\sigma}^2)$ or of $(\hat{\mu}, s^2)$?
- It turns out that $\bar{X}$ and $\sum_{i=1}^n (X_i - \bar{X})^2$ are independent.
- Therefore, $\hat{\mu}$ and s^2 are independent. (So are $\hat{\mu}$ and $\hat{\sigma}^2$.)

A heuristic outline of the proof (See Section 8.3 in textbook for a more detailed proof)

- Suppose $Z_1, Z_2, \ldots, Z_n$ are i.i.d. $N(0, 1)$ random variables.
- Let $(Z_1, Z_2, \ldots, Z_n)$ be the coordinate of a random point in $\mathbb{R}^n$.
- Note that the joint p.d.f of $(Z_1, Z_2, \ldots, Z_n)$ depends on the data through $\sum_{i=1}^{n} Z_i^2$—the squared distance of the point from the origin.
- (Draw a figure.)

- By symmetry, if one arbitrarily "rotates" the orthogonal axes, the new coordinates will have the same joint distribution, that is, they will still be i.i.d. $N(0, 1)$.
- Now rotate the axis such that one of them lie on $(1, 1, \ldots, 1)$.
- Then the coordinate along that axis is $\frac{\sum_{i=1}^{n} Z_i}{\sqrt{n}}$.
- The squared sum of the other $(n-1)$ new coordinates are

$$\sum_{i=1}^{n} Z_i^2 - \left(\frac{\sum_{i=1}^{n} Z_i}{\sqrt{n}}\right)^2 = \sum_{i=1}^{n} (Z_i - \bar{Z})^2.$$

- Therefore $\sum_{i=1}^{n} Z_i$ is independent of $\sum_{i=1}^{n} (Z_i - \bar{Z})^2$, and

$$\sum_{i=1}^{n} (Z_i - \bar{Z})^2 \sim \chi^2(n-1).$$

- Finally, if $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$.
- We can let $Z_i = \frac{X_i - \mu}{\sigma}$ for $i = 1, 2, \ldots, n$.
- Now note that

$$\bar{X} = \mu + \sigma \bar{Z} \quad \text{and} \quad \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^n (Z_i - \bar{Z})^2.$$

- This completes the proof.

A brain teaser ...

- Two statisticians are both trying to estimate the mean of a normal distribution $N(\mu, \sigma^2)$.
- Each of them are basing their estimate on 10 independent observations from the *same* distribution.
- The 10 observations that Statistican 1 have are different from the what Statistician 2 had observed.
- For Statistician 1, the sample average $\bar{X}$ is 10 and the sample standard deviation s is 8.
- For statistician 2, the sample average $\bar{X}$ is 15 and the sample standard deviation s is 6.
- They both want to use $\bar{X}$ to estimate μ.
- *Whose estimate for μ is more trustworthy?*

Approximate sampling distribution for estimators

- Up till now we have been focusing on finding the exact sampling distributions for the MLE.
- Such cases are actually not common—in most cases the sampling distribution of MLE cannot be calculated explicitly.
- Instead, most often we have to rely on approximation to the sampling distribution.
  - Theoretical approximation vs simulation-based approximation.
- The *Central Limit Theorem* (CLT) provides the basis for such approximation in a broad range of applications.

The Central Limit Theorem (A heuristic)

If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with

$$E(X_i) = \mu \quad \text{and} \quad \mathrm{Var}(X_i) = \sigma^2 < \infty,$$

then if n is large,

$$\sum_{i=1}^{n} X_i \quad \text{is approximately } N(n\mu, n\sigma^2)$$

and thus

$$\bar{X} \quad \text{is approximately } N\left(\mu, \frac{\sigma^2}{n}\right).$$

Therefore when we have a lot of data, the sample mean is distributed "as if" the data came out of the normal distribution.

χ^2 distribution with n degrees of freedom

If

$$Y = U_1^2 + U_2^2 + \dots U_n^2,$$

where $U_1^2, U_2^2, \dots U_n$ are i.i.d χ_1^2 random variables. In particular

$$E(U_1^2) = 1 \quad \text{and} \quad \mathrm{Var}(U_1^2) = 2.$$

So when n is large, $Y = \sum_{i=1}^n U_i^2$ is approximately $N(n, 2n)$.

The χ^2 distribution with k degrees of freedom

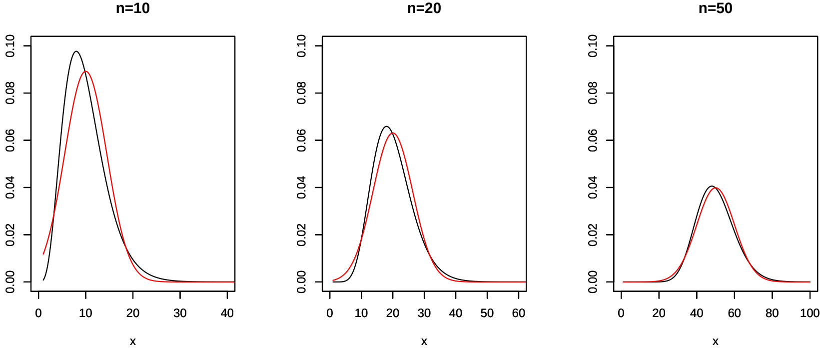

Figure: Red curves are pdfs for $N(n, 2n)$; black curvess are pdfs for $\chi^2(n)$.

- For large n is the pdf of χ_n^2 and that for $N(n, 2n)$ is almost the same, although for $n = 1$ the two distributions are very different!

How large n needs to be before we can apply the theorem to $\sum_{i=1}^{n} X_i$?

- The answer depends on the distribution of the $X_i^{\prime}s$.
- If that distribution is very different from normal than n needs to be quite large. *A rough rule of thumb is $n \geq 30$.*
- If that distribution is similar to normal to begin with, then the approximation can be very good with n as small as 5.

The CLT (formal version)

More formally, the theorem is stated in *standardized* form:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \quad \text{is approximately } N(0,1),$$

and similarly,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{is approximately } N(0,1).$$

Example: One use of this approximate sampling distribution

- For any two constants a and b such that $a < b$, what is the probability for $\bar{X}$ to be in the range (a, b). Now when n is large,

$$P(a < \bar{X} < b) = P\left(\frac{a-\mu}{\sigma/\sqrt{n}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{b-\mu}{\sigma/\sqrt{n}}\right).$$

Let $W = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. By the CLT, we know W is distributed approximately as a $N(0,1)$ random variable. Therefore

$$\begin{aligned} P(a < \bar{X} < b) &= P\left(\frac{a-\mu}{\sigma/\sqrt{n}} < W < \frac{b-\mu}{\sigma/\sqrt{n}}\right) \\ &\approx \Phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right). \end{aligned}$$

For example, we are often interested in calculating the probability for the estimator $\bar{X}$ to “miss” μ by an amount less than some constant $\varepsilon > 0$.

$$P(|\bar{X} - \mu| < \varepsilon) = P(\mu - \varepsilon < \bar{X} < \mu + \varepsilon),$$

so we are in the previous scenario where $a = \mu - \varepsilon$ and $b = \mu + \varepsilon$.

Therefore

$$
\begin{aligned}
P(|\bar{X}-\mu|<\varepsilon) &\approx \Phi\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right)-\Phi\left(-\frac{\varepsilon}{\sigma/\sqrt{n}}\right)\\
&=\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right)-\Phi\left(-\frac{\sqrt{n}\varepsilon}{\sigma}\right)\\
&=\int_{-\frac{\sqrt{n}\varepsilon}{\sigma}}^{\frac{\sqrt{n}\varepsilon}{\sigma}}\phi(x)dx.
\end{aligned}
$$

Note that in particular as n becomes very large, this probability gets close to

$$\int_{-\infty}^{\infty}\phi(x)dx=1.$$

This result is called the *(weak) law of large number*. It is very intuitive: as you get more and more data,

the sample average "converges" to the actual population mean.

Coming up next . . .

- Finding (approximate) sampling distributions for MLEs.
- Fisher's approximation.