STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 10

The sample variance

- Last time we see that for i.i.d. data from a normal distribution with unknown mean and variance, the estimator

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  is the MLE.
- We see that this is biased, as its expectation is

$$\begin{aligned} E(\widehat{\sigma^2}) &= \frac{1}{n}E\left(\sum_{i=1}^{n}X_i^2 - \frac{(\sum_{i=1}^{n}X_i)^2}{n}\right) \\ &= \frac{1}{n}\left(n\sigma^2 + n\mu^2 - \frac{n\sigma^2 + n^2\mu^2}{n}\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2. \end{aligned}$$

- An unbiased estimator is thus the *sample variance*

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

- If the data are i.i.d. from some other distribution other than normal, the above calculation of expectation still holds, provided that $E(X^2)$, or equivalently $\mathrm{Var}(X) = \sigma^2 < \infty$.
- Thus for non-normal data with unknown mean and variance, these two statistics are still meaningful estimators for the variance, and we still have

$$E(\widehat{\sigma^2}) = \frac{n-1}{n}\sigma^2 \quad \text{and} \quad E(s^2) = \sigma^2.$$

- The difference is that without normality $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is no longer the MLE.
- *Question: Why does $\widehat{\sigma^2}$ underestimate the actual variance σ^2?*

The sampling distributions of estimators

- Sampling distributions of estimators characterize how well the estimators perform for different values of the parameters.
- The mean, the variance, and the MSE are summary statistics for the sampling distributions.
- They can often be calculated without knowing the exact form of the sampling distributions.
- However sometimes it will be useful to know the actual sampling distribution.
  - For example, to compute $P_\theta(|\hat{\theta} - \theta| < c)$ for any c.
  - For another example, to construct interval estimates as we will see later.

Finding the sampling distributions of common estimators

- In some situations, the sampling distribution of an estimator is easy to find.
- For instance, in the political poll example, the distribution of the MLE for θ, $\hat{\theta} = X/n$ can be found with a simple change of variable:

$$P(\hat{\theta} = u|\theta) = \binom{n}{nu} \theta^{nu}(1-\theta)^{n(1-u)} \quad \text{for } u = 0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1.$$

  (Exercise: What's the sampling distribution for $\delta(X) = \frac{X+1}{n+2}$?)
- More generally, finding the sampling distribution is not as straightforward.
- We next cover several techniques useful for this purpose.

The distribution of sums of independent random variables

- Many common estimators are sums of independent random variables.
- For example, the MLE for the mean θ of an exponential distribution (e.g. lifetime of light bulbs), or a normal distribution (e.g. actual amount of air pollutant), based on i.i.d. observations is $\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- A useful fact for finding the distribution of "sums" is the following.

Let (X, Y) be a bivariate random variable with density $f(x, y)$, then the density of

$$Z = X + Y$$

is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f(z-y, y)dy.$$

If X and Y are independent then $f(x, y) = f_X(x)f_Y(y)$, and so

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy.$$

Scratch of the proof

To find the p.d.f of Z, let us start with its c.d.f:

$$F_Z(z) = P(Z \leq z) = P(X+Y \leq z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f(x,y)dxdy.$$

(Draw a figure.)

Therefore

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} \frac{d}{dz} \left(\int_{-\infty}^{z-y} f(x,y)dx \right) dy = \int_{-\infty}^{\infty} f(z-y,y)dy.$$

A "physics" proof.

Example: The sum of independent normal random variables

- Let X and Y be two independent normal random variables.
- Suppose $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\theta, \tau^2)$.
- Then the p.d.f of $Z = X + Y$ is

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-y-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{(y-\theta)^2}{2\tau^2}} dy \\ &= \frac{1}{2\pi\sigma\tau} \int_{-\infty}^{\infty} e^{-\frac{1}{2}[(z-y-\mu)^2/\sigma^2 + (y-\theta)^2/\tau^2]} dy. \end{aligned}$$

Note that

$$(z-y-\mu)^2/\sigma^2 + (y-\theta)^2/\tau^2$$

is a quadratic form in y and z, so after completing some squares, it can be written as

$$A(z-B)^2 + C(y-Dz+E)^2 + F.$$

Evaluating A, B, C, D, E and F is straightforward but tedious.

- Fortunately, we don't have to evaluate them. We can rewrite

$$\begin{aligned} f_Z(z) &= \frac{1}{\sqrt{2\pi}\sigma\tau\sqrt{C}} e^{-\frac{F}{2}} \cdot e^{-\frac{A}{2}(z-B)^2} \cdot \int_{-\infty}^{\infty} \sqrt{\frac{C}{2\pi}} e^{-\frac{C}{2}(y-Dz+E)^2} dy \\ &= \frac{1}{\sqrt{2\pi}\sigma\tau\sqrt{C}} e^{-\frac{F}{2}} \cdot e^{-\frac{A}{2}(z-B)^2} \cdot 1 \\ &\propto e^{-\frac{A}{2}(z-B)^2}. \end{aligned}$$

- So we can tell that Z must have a normal distribution with mean B and variance $1/A$!
- Now

$$\begin{aligned} E(Z) &= E(X) + E(Y) = \mu + \theta \\ \mathrm{Var}(Z) &= \mathrm{Var}(X) + \mathrm{Var}(Y) = \sigma^2 + \tau^2. \end{aligned}$$

- Hence Z has a $N(\mu+\theta, \sigma^2+\tau^2)$ distribution.

- By induction, we can then show that if $X_1, X_2, \ldots, X_n$ are independent random variables, with $X_i \sim N(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

- In particular, if the X_i's are i.i.d. $N(\mu, \sigma^2)$, then

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

and

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

- Now we can calculate quantities such as $P(|\bar{X} - \mu| < c)$ for any c based on this distribution.

The χ^2 distribution with 1 degree of freedom

- If $U \sim N(0,1)$, then we call the distribution of U^2, the

  χ^2 (or Chi-square) distribution with 1 degree of freedom,

- This is often denoted as the $\chi^2(1)$, or χ^2_1 distribution.
- Its p.d.f can be found by a change-of-variable (note though the mapping is not one-to-one)

$$f_{U^2}(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2} \quad \text{for } y > 0$$
$$= 0 \quad \text{otherwise.}$$

Note that this is just the density of Gamma(1/2,1/2).

- Now if we have n i.i.d. $N(0,1)$ random variables $U_1, U_2, \ldots, U_n$, then the probability distribution of their sum

$$Z = U_1^2 + U_2^2 + \cdots + U_n^2$$

is called the *χ^2 (or Chi-square) distribution with n degrees of freedom*, denoted by $\chi^2(n)$ or χ^2_n.

- We can think of the number of *degrees of freedom* (d.f.) as the number of "free" (independent) pieces that constitute the sum.
- The p.d.f of Z is

$$\begin{aligned} f_{\chi^2_n}(z) &= \frac{1}{2^{n/2}\Gamma(n/2)} z^{\frac{n}{2}-1} e^{-\frac{z}{2}} \quad \text{for } z > 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

We can verify this p.d.f again using induction.

- First, check that when $n = 1$, since $\Gamma(n/2) = \sqrt{\pi}, f_{\chi^2(1)}(z)$ takes the above form.
- Now let

$$X = U_1^2 + U_2^2 + \cdots + U_{k-1}^2$$

and

$$Y = U_k^2.$$

- From the induction hypothesis,

$$\begin{aligned} f_X(x) &= \frac{1}{2^{\frac{k-1}{2}}\Gamma\left(\frac{k-1}{2}\right)} x^{\frac{k-1}{2}-1} e^{-\frac{x}{2}} \quad \text{for } x > 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

and Y has the p.d.f for $\chi^2(1)$.

Therefore

$$
\begin{aligned}
f_{\chi^2(k)}(z) &= \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy \\
&= \int_0^z \frac{1}{2^{\frac{k-1}{2}} \Gamma\left(\frac{k-1}{2}\right)} \cdot (z-y)^{\frac{k-1}{2}-1} \cdot e^{-\frac{z-y}{2}} \cdot \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}} dy \\
&= \frac{1}{2^{\frac{k-1}{2}} \Gamma\left(\frac{k-1}{2}\right) \sqrt{2\pi}} e^{-\frac{z}{2}} \int_0^z (z-y)^{\frac{k-3}{2}} y^{-\frac{1}{2}} dy.
\end{aligned}
$$

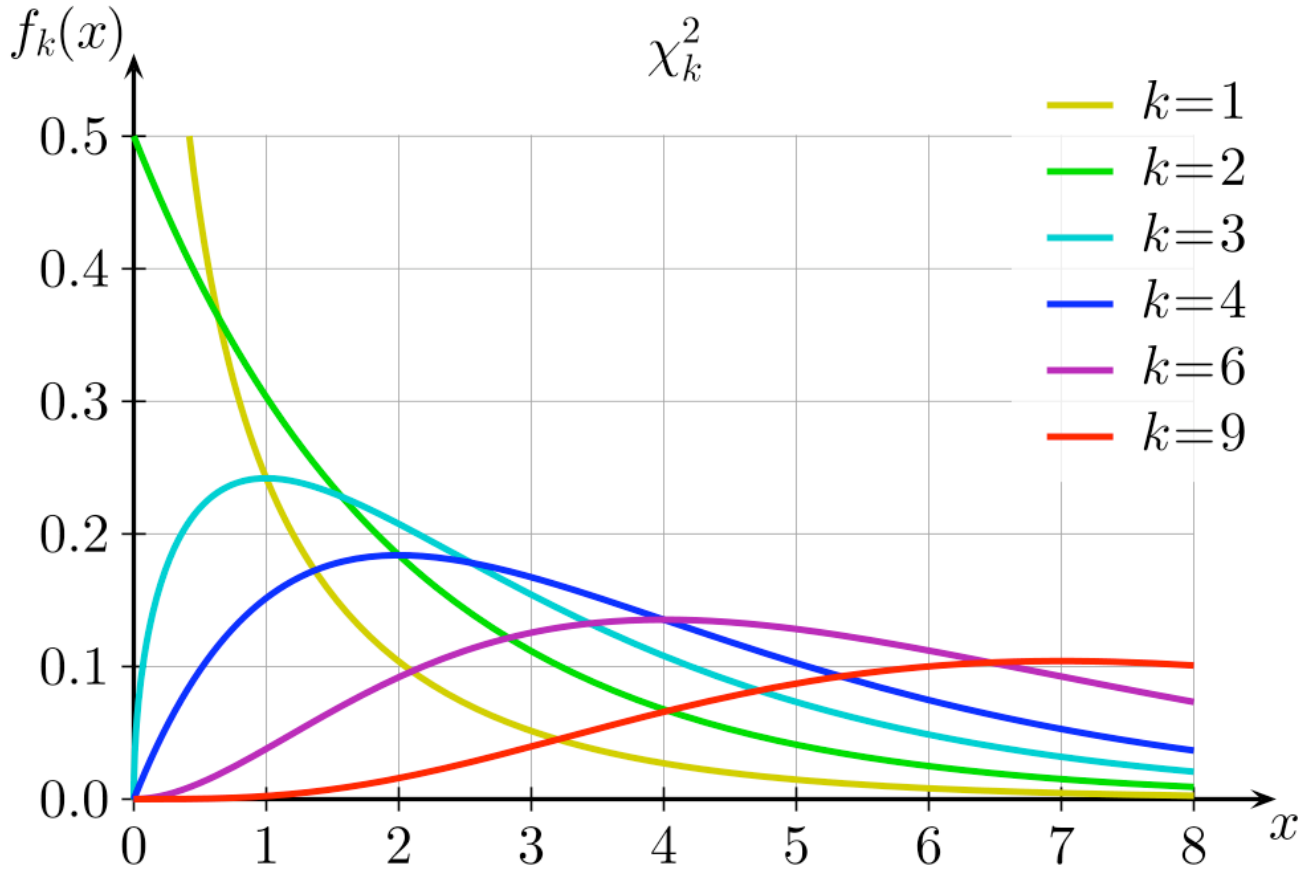
Now let $u = y/z$, by a change of variable (check!)

$$
\begin{aligned}
\int_0^z (z-y)^{\frac{k-3}{2}} y^{-\frac{1}{2}} dy &= z^{\frac{k-3}{2}} \cdot z^{-\frac{1}{2}} \cdot z \int_0^1 (1-u)^{\frac{k-3}{2}} u^{-\frac{1}{2}} du \\
&= z^{\frac{k}{2}-1} \cdot \frac{\Gamma(\frac{1}{2})\Gamma(\frac{k-1}{2})}{\Gamma(\frac{k}{2})}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
f_{\chi^2(k)}(z) &= \frac{1}{2^{k/2}\Gamma(k/2)} z^{\frac{k}{2}-1} e^{-\frac{z}{2}} \quad \text{for } z > 0 \\
&= 0 \quad \text{otherwise.}
\end{aligned}
$$

The χ^2 distribution with k degrees of freedom



Source: Wikipedia

One important use of the χ^2 distribution

- It turns out that if $X_1, X_2, \ldots, X_n$ are i.i.d. $N(0,1)$ random variables, then
$$\sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2_{n-1}.$$
- Hence (how?), if $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables, then
$$\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}.$$
Note that
$$\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} = \frac{n\widehat{\sigma^2}}{\sigma^2}.$$
- This gives the sampling distribution of s^2 and $\widehat{\sigma^2}$.

Example

Let $X_1, X_2, \ldots, X_n$ be i.i.d data from $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. We use the sample variance to estimate σ^2.

- What is the probability for s^2 to make a relative error no more than 2%?
- That is, what is $P(|s^2 - \sigma^2| < 0.02\sigma^2)$?

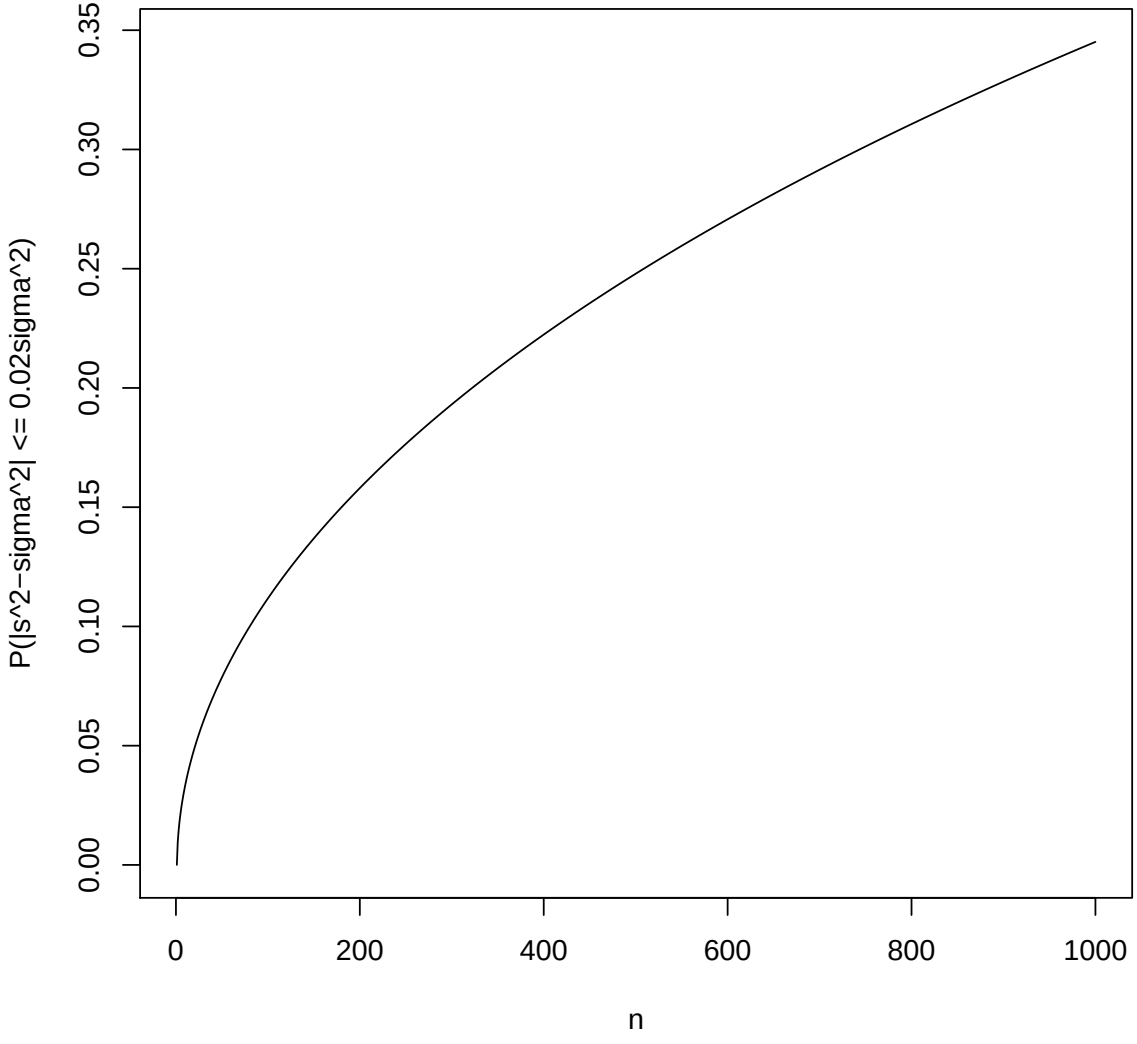
To calculate this probability, we note that

$$\begin{aligned} P(|s^2 - \sigma^2| \leq 0.02\sigma^2) &= P\left(0.98 \leq \frac{s^2}{\sigma^2} \leq 1.02\right) \\ &= P\left(0.98(n-1) \leq \frac{(n-1)s^2}{\sigma^2} \leq 1.02(n-1)\right) \\ &= F_{\chi^2_{n-1}}\big(1.02(n-1)\big) - F_{\chi^2_{n-1}}\big(0.98(n-1)\big). \end{aligned}$$

So if $n = 100$, this is

$$F_{\chi^2_{99}}(100.98) - F_{\chi^2_{99}}(97.02) \approx 0.574 - 0.462 = 0.112.$$

Sample size vs $P(|s^2 - \sigma^2| < 0.02\sigma^2)$?



Properties of the χ^2 distribution

- If X and Y are independent random variables with

$$X \sim \chi^2_n \quad \text{and} \quad Y \sim \chi^2_m,$$

then

$$X + Y \sim \chi^2(n+m).$$

So the d.f. of independent χ^2 random variables add up. Why?

A heuristic proof

If we represent X and Y as

$$X = U_1^2 + U_2^2 + \cdots + U_n^2$$

and

$$Y = U_{n+1}^2 + U_{n+2}^2 + \cdots + U_{n+m}^2.$$

Then

$$X + Y = U_1^2 + U_2^2 + \cdots + U_{n+m}^2 \sim \chi^2(n+m).$$

The mean and variance of χ^2_n

Let $Z \sim \chi^2_n$. By its representation as the sum of independent random variables

$$Z = U_1^2 + U_2^2 + \cdots U_n^2$$

with $U_i \sim N(0,1)$, we have

$$E(Z) = E(\sum_{i=1}^{n} U_i^2) = nE(U_1^2) = n.$$

and

$$\mathrm{Var}(Z) = \mathrm{Var}\left(\sum_{i=1}^{n} U_i^2\right) = n\mathrm{Var}(U_1^2).$$

Now let us find $\mathrm{Var}(U_1^2)$. First, let $Y = U_1^2$. Then

$$E(Y) = \mathrm{Var}(U_1) + E(U_1)^2 = \mathrm{Var}(U_1) = 1.$$

and

$$E(Y^2) = \int_0^\infty y^2 \cdot \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}} dy = \frac{1}{\sqrt{2\pi}} \int_0^\infty y^{\frac{3}{2}} e^{-\frac{y}{2}} dy.$$

Let $t = y/2$, we get

$$\begin{aligned} E(Y^2) &= \frac{2^{\frac{5}{2}}}{\sqrt{2\pi}} \int_0^\infty t^{\frac{5}{2}-1} e^{-t} dt \\ &= \frac{2^{\frac{5}{2}}\Gamma(\frac{5}{2})}{\sqrt{2\pi}} = 4 \cdot \frac{\frac{3}{2} \cdot \frac{1}{2} \cdot \Gamma(\frac{1}{2})}{\sqrt{\pi}} = 3. \end{aligned}$$

where we used two facts:

- $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Therefore

$$\mathrm{Var}(U_1^2) = \mathrm{Var}(Y) = E(Y^2) - (E(Y))^2 = 3 - 1 = 2,$$

and thus

$$\mathrm{Var}(Z) = 2n.$$

- In your homework, you will need to use this to find the variance of s^2 and $\widehat{\sigma^2}$ for normal i.i.d. data.
- For example,

$$E(\widehat{\sigma^2}) = \frac{\sigma^2}{n} E\left(\frac{n\widehat{\sigma^2}}{\sigma^2}\right) = \frac{\sigma^2}{n}(n-1) = \frac{n-1}{n}\sigma^2.$$

Connection with Gamma distribution

- Recall that a random variable with the Gamma(α, β) distribution has p.d.f

$$f(x|\alpha,\beta) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad \text{for } x > 0$$
$$= 0 \quad \text{otherwise.}$$

- So the χ^2_n distribution is exactly Gamma$(n/2, 1/2)$.
- Gamma(α, β) has mean α/β and variance α/β^2.
- Verify that this gives the same mean and variance for χ^2_n with $\alpha = n/2$ and $\beta = 1/2$.

- 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 $\hat{\sigma}^2$.)