STAT 315: Sampling Distributions

Luc Rey-Bellet

University of Massachusetts Amherst

luc@math.umass.edu

April 24, 2025

Independent Identically Distributed (=IID) Random Variables

Operational Meaning: Repeat an experiment n times, each time under the exactly same conditions, each time independently of the other experiments. The outcome of each experiment is a measurement Y_i, $i = 1, 2, \cdots, n$ (a random variable).

IID RV and sample mean

- $Y1, \cdots, Y_n$ are n independent RV.
- $Y1, \cdots, Y_n$ are identically distributed: The PDF of each Y_i is $f(y_i)$ with the same f.
- The joint PDF is $f(y_1, \cdots, y_n) = f(y_1) \cdots f(y_n)$.
- The mean $E[Y_i] = \mu$ and the variance $V[Y_i] = \sigma^2$ do no depend of i.
- The sample mean is

$$\overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

The Law of Large Numbers

Remember

Mean and Variance

If $Y_1, \cdots Y_n$ IID with $E[Y_i] = \mu$ and $V[Y_i] = \sigma^2$ then

$$E[\overline{Y}] = E\left[\frac{Y_1 + \cdots Y_n}{n}\right] = \mu \qquad V[\overline{Y}] = V\left[\frac{Y_1 + \cdots Y_n}{n}\right] = \frac{\sigma^2}{n}$$

So by Chebyshev Inequality

Law of Large Numbers

If $Y_1, \cdots Y_n$ IID with $E[Y_1] = \mu$ and $V[Y_1] = \sigma^2$ then

$$P\left\{|\overline{Y} - \mu| > \epsilon\right\} \leq \frac{V[\overline{Y}]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \underbrace{\rightarrow}_{\text{as } n \to \infty} 0$$

Interpretation of the Law of Large Numbers: Concentration of the sample mean $\overline{Y}$ around the true mean μ

The variance of $\overline{Y}$ decreases with n so the PDF of $\overline{Y}$ is more and more concentrated around the mean μ.

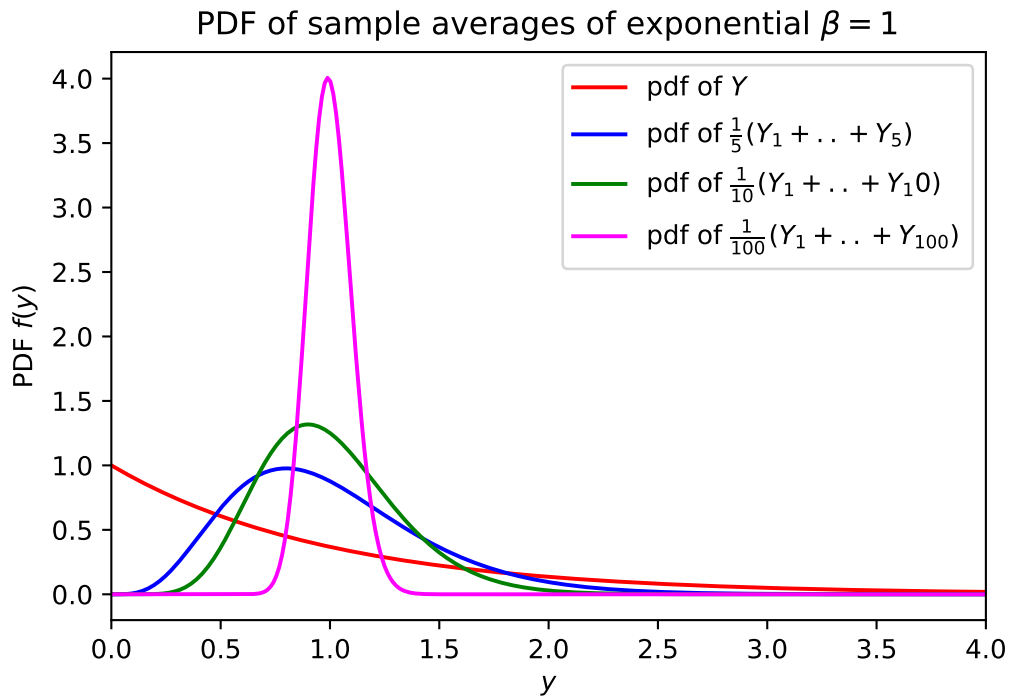

Figure: If Y_i is exponential with parameter 1 then $\frac{Y_1+\cdots+Y_n}{n}$ is gamma with $\alpha = 1$ and $\beta = \frac{1}{n}$ see previous slides.

Using the Law of Large numbers to estimate an unknown μ

Operational meaning: You can perform an experiment resulting a measurement Y but you do not know the PDF of Y neither do you know the mean μ? The Law of large numbers tells you to perform n independent

experiment $Y_1, Y_2, \cdots, Y_n$ and use the approximation

$$\frac{Y_1 + \cdots + Y_n}{n} \approx \mu$$

Chebyshev Theorem tells you what this is a good approximation with high probability as n grows.

$\longrightarrow$ Can we figure out how good this approximation really is? The central limit theorem provides one answer: all sample mean looks like normal random variables for large n.

The Monte-Carlo method

Put the law of large numbers to work: Imagine μ is some quantity you try to evaluate.

- Find a random variable X with $E[X] = \mu$.
- Simulate $X_1, X_2, \cdots, X_n$ on a computer.
- Get the estimate $\mu = \frac{X_1 + \cdots + X_n}{n}$

Example: Estimate the number π using random numbers

- Let U and V be two random numbers. If $U^2 + V^2 \leq 1$ set $Y = 1$, otherwise set $X = 0$. Then

$$P(Y = 1) = P(U^2 + V^2 \leq 1) = \frac{\pi}{4}$$

  and $E[Y] = \frac{\pi}{4}$.
- Taking IID copies $Y_1, X_2, , Y_n$ we have

$$4\frac{Y_1 + \cdots + Y_n}{n} \approx 4E[Y]$$

```python
import random

def monte_carlo_pi(num_points):
    circle_points = 0
    total_points = num_points

    for _ in range(num_points):
        x = random.uniform(-1, 1)
        y = random.uniform(-1, 1)

        # Check if the point is within the unit circle
        if x**2 + y**2 <= 1:
            circle_points += 1

    # Estimate pi
    pi_estimate = 4 * circle_points / total_points
    return pi_estimate

# Number of points for the estimation
num_points = 100000

# Estimate pi
pi_estimate = monte_carlo_pi(num_points)
print("Estimated value of π:", pi_estimate)
```

PDF of the sample mean $\overline{Y}$

Except in special cases it is very difficult to compute the PDF of $\overline{Y}$.

Example: Take $n = 3$ and toll three dice. Y_i =number on dice i.

$$\overline{Y} = \frac{Y_1 + Y_2 + Y_3}{3}.$$

Since $Y_1 + Y_2 + Y_3$ takes (integer) values between 3 and 18 the sample average $\overline{Y}$ takes value in

$$\left\{1, \frac{4}{3}, \frac{5}{3}, \frac{6}{3}, \cdots, \frac{18}{3} = 6\right\}$$

.

$$\begin{aligned}
P(\overline{Y} = 1) &= p(1,1,1) = \frac{1}{216} \\
P(\overline{Y} = 4/3) &= p(1,1,2) + p(1,2,1) + p(2,1,1) = \frac{3}{216} \\
P(\overline{Y} = 5/3) &= p(1,1,3) + p(1,3,1) + p(3,1,1) \\
&\quad + p(2,2,1) + p(1,2,2) + p(2,1,2) = \frac{6}{216}
\end{aligned}$$

Sample mean for normal random variables

For normal random variables everything can be computed explicitly so let us start here. The Central limit theorem will tell us that everything is normal..

Use the fact that if Y_1 is normal with mean μ_1 and variance σ_1^2 and Y_2 is normal with mean μ_2 and variance σ_2^2 then

$a_1 Y_1 + a_2 Y_2$ is normal with mean $a_1\mu_1 + a_2\mu_2$ and variance $a_1^2\sigma_1^2 + a_2^2\sigma_2^2$

So if s Y_i normal with mean μ and variance σ^2:

$$\overline{Y} = \frac{Y_1 + \cdots Y_n}{n} \text{ is normal with mean } \mu \text{ and variance } \sigma^2/n$$

Example: estimating an unknown μ.

A bottling machine fills bottles with a normal distribution unknown(!) mean and a standard deviation of $\sigma = 1$ fl. oz.
If you fill 9 bottles what is the probability that the mean μ is within .3 fl.oz of the sample mean $\overline{Y}$?

Standardize with the variance $\frac{\sigma^2}{n}$.

$$\begin{aligned} P(|\overline{Y} - \mu| \leq 0.3\} &= P(-0.3 \leq \overline{Y} - \mu \leq 0.3) \\ &= P\left(\frac{-0.3}{\sigma/\sqrt{n}} \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.3}{\sigma/\sqrt{n}}\right) \\ &= P\left(\frac{-0.3}{1/\sqrt{9}} \leq Z \leq \frac{0.3}{1/\sqrt{9}}\right) \quad Z \text{ standard normal} \\ &= P(-.9 \leq Z \leq .9) = .6318 \quad \text{Use z-score table} \end{aligned}$$

Say if we observe 9 bottles with an average of 19.5 fl. oz then the true mean μs is in the interval $[19.2, 19.8]$ with probability .6318?

Example: estimating an unknown μ, continued

How many bottles should you fill for $\overline{Y}$ to be no more than .3 ounces from μ with probability .95?

Use that for Z a standard normal we have

$$P(-.1.96 \leq Z \leq 1.96) = .95 \text{Use z-score table.}$$

$$\begin{aligned} P\left(|\overline{Y} - \mu| \leq .3\right) &= P\left(\frac{-0.3}{\sigma/\sqrt{n}} \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.3}{\sigma/\sqrt{n}}\right) \\ &= P\left(\frac{-0.3}{1/\sqrt{n}} \leq Z \leq \frac{0.3}{1/\sqrt{n}}\right) \\ &= P\left(-0.3\sqrt{n} \leq Z \leq 0.3\sqrt{n}\right) \end{aligned}$$

And thus

$$.3\sqrt{n} = 1.96 \Leftrightarrow n = \left(\frac{1.96}{.3}\right)^2 = 42.68$$

We need 43 bottles for a 95% confidence interval of .3 fl. oz

Sample Variance: what if we don't know σ^2.

- If μ is unknown we can estimate μ using sample averages $\bar{Y}$
- How can we estimate σ^2 from the sample $Y_1, Y_2, \cdots, Y_n$?

First attempt: μ is known and we want σ^2. (Not realistic but...)
We use that

$$V[Y] = E[(Y-\mu)^2] = E[Z] \quad \text{with} \quad Z = (Y-\mu)^2$$

The sample mean for Z is

$$\overline{Z} = \frac{Z_1 + Z_2 + \cdots Z_n}{n} = \frac{(Y_1-\mu)^2 + \cdots (Y_n-\mu)^2}{n} \approx E[Z] = \sigma^2$$

by the Law of Large Numbers applied to $Z_i = (Y_i - \mu)^2$

Sample Variance, continued

If we do not know μ we replace μ by the sample mean $\overline{Y}$.

Sample Variance

The sample variance is defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{(Y_1 - \overline{Y})^2 + \cdots (Y_n - \overline{Y})^2}{n-1}$$

Note the factor $n-1$ in the denominator. We have

$$P(|S^2 - \sigma^2| \geq \epsilon) \to 0 \text{ as } n \to \infty$$

and thus $S^2 \approx \sigma^2$ for large n with high probability and

$$E[S^2] = \sigma^2$$

for any n, i.e S^2 is an unbiased estimator.

Why the factor $n-1$?

- If n is large using $1/n$ or $1/(n-1)$ does not matter much in practice
- The factor $1/(n-1)$ ensures that $E[S^2] = \sigma^2$ (unbiased estimator)
  This can be shown by a somewhat long computation....

Sample variance for normal random variables

Recall some facts about normal and gamma (and χ^2) RV.

- If Z standard normal then Z^2 is gamma with $\alpha = 1/2$ and $\beta = 2$ (also called χ^2).
- X_1 gamma with α_1 and β and X_2 gamma with α_2 and β then $X_1 + X_2$ is gamma with $\alpha_1 + \alpha_2$ and β then
- If X is gamma with α and β then $Y = aX$ gamma with α and $a\beta$ then

So if Y_i are independent normal with mean μ and variance σ^2 then

$$\left(\frac{Y_i - \mu}{\sigma}\right)^2 \text{ is a } \chi^2, \text{ aka gamma with } \alpha = 1/2, \beta = 2$$

$$\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2 \text{ is a } \chi^2_n, \text{ or gamma with } \alpha = n/2, \beta = 2$$

$$\frac{1}{n}\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2 \text{ is gamma with } \alpha = n/2, \beta = 2/na$$

Confidence interval for chi^2 RV

Example: Find a 95% confidence interval for $\sum_{i=1}^{6} \left(\frac{Y_i-\mu}{\sigma}\right)^2$.

Since $Z_i = \frac{Y_i-\mu}{\sigma}$ are IID standard normal RV then $\sum_{i=1}^{6} Z_i^2$ is gamma with $\alpha = 3$ and $\beta = 2$. So in particular $E[Y] = 6$.
To find a confidence interval we need to find

$$b \text{ such that } P(Y \geq b) = .025$$

$$a \text{ such that } P(Y \leq a) = .025$$

Use for example the online calculator
`https://homepage.divms.uiowa.edu/~mbognar/`
one finds $b = 14.44$ and $a = 1.23$

$$P\left(1.23 \leq \sum_{i=1}^{6} \left(\frac{Y_i - \mu}{\sigma}\right)^2 \leq 14.44\right) = .95$$

Sample Variance for normal

If we replace μ by $\overline{Y}$ one can show (more long computations) that

Sample variance for normal random variable

Suppose Y_i are independent normal with mean μ and variance σ^2. Then

$$\sum_{i=1}^{n} \left(\frac{Y_i - \overline{Y}}{\sigma} \right)^2 \text{ is } \chi^2_{n-1} \text{ or gamma with } \alpha = (n-1)/2, \beta = 2$$

and the $n-1$ factor is here again!

Therefore

$$\frac{S^2}{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{Y_i - \overline{Y}}{\sigma} \right)^2 \text{ is gamma with } \alpha = (n-1)/2, \beta = 2/n-1$$

95% confidence interval for the variance

.

Example: Coming back to our machine bottling example with now unknown μ and σ^2.

Suppose we fill 50 bottles and measure the values $Y_1 = 19.5$ $Y_2 = 19.7$ ... $Y_{50} = 20.1$ which gives a value

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = .28$$

We know that $\frac{S^2}{\sigma^2}$ is gamma with $\alpha = \frac{50-1}{2} = 24.5$ and $\beta = \frac{2}{50-1} = \frac{1}{24.5}$ and so using the online calculator we find

$$P\left(.64 \leq \frac{S^2}{\sigma^2} \leq 1.43\right) = .95$$

and with our experimental value $S^2 = .28$

$$P\left(\frac{S^2}{1.23} \leq \sigma^2 \leq \frac{S^2}{.64}\right) = P\left(.19 \leq \sigma^2 \leq .43\right) = .95$$