Topic II: Properties of a Random Sample

Wei You

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Fall, 2023

Population and Sample

- A *population* is a set of similar objects that we want to understand its properties.
- The first step of statistical research is usually to collect a data sample from a population by some procedure. The statistical property of the sample tells us about the population.
- *A complete sample:* a sample that includes ALL objects satisfying some selection criteria. It is infeasible most of the times.
- A compromise is to use *an unbiased (representative) sample:* a sample that does not depend on the properties of the objects.

Selection Bias

- Inspection paradox: how should we estimate the length-of-stay of the tourist at the Disneyland resort?
- Survivorship bias.
- Simpson's paradox.
- ...

Random Sample

One simplification is to deal with the random samples.

Random Sample

If $X_1, X_2, \ldots, X_n$ are independent random variables having a common distribution F (a.k.a. i.i.d.), then we say that they constitute a random sample of size n from the distribution F.

For some parameter θ, the joint PMF/PDF is

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Example: Let $X_1, \ldots, X_n$ be a random sample from an exponential(β) population. What is the joint PDF/CDF of the sample?

Sample from a Finite Population

- In reality, even for a large (finite) population, the distributions of X_1 and X_2 are not exactly i.i.d.; it is sampled *without replacement*.
- What if the population size is small?
- **Example:** There are 100 balls, numbering 1 to 100; we sample 10 from them.
- Sampling from a finite population without replacement is called simple random sampling. It is different from the random sample definition.
- But don't worry, when the population is not too small, the difference is small.
- **Example:** Suppose we have a population of size 1000: $\{1, \ldots, 1000\}$. We sample 10 from them without replacement. What is $\mathbb{P}(X_1 > 200, \ldots, X_{10} > 200)$? 0.106164 vs 0.107374 (with replacement).
- We focus on the i.i.d. random sample throughout the course.

Sample Statistics

Statistic

A statistic $T(X_1, \ldots, X_n)$ is a random variable whose value is determined by the sample.

It is not a function of the parameter! Suppose we have a sample from a population with distribution $X_1, X_2, \ldots, X_n \sim F$.

Sample mean and variance

The sample mean is defined by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

The sample variance S^2 is defined by

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

$S = \sqrt{S^2}$ is called the sample standard deviation. For sample size one, S is not defined.

The Ultimate Question

Are They Useful?

What can $\bar{X}$ and S^2 tell us about μ and σ^2 of F?

Sample Mean

The properties of the sample mean.

- Unbiasedness: the expected value of a statistic is equal to the parameter it intends to estimate.

Unbiasedness

Sample mean is unbiased $E[\bar{X}] = E\left[\frac{X_1+\cdots+X_n}{n}\right] = \frac{1}{n}(E[X_1] + \cdots + E[X_n]) = \mu$

- If F has finite variance σ^2, then

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2}[\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

As n increases, the sample mean becomes less and less random.

Sample Mean minimizes Mean Squared Error (MSE)

$\bar{X}$ minimizes the MSE

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

So the sample mean is a "center" of the sample.

Connection to the MSE of estimating a random variable Y using a single number c.

An Algebraic Identity

If $\bar{x} = \sum_{i=1}^{n} x_i/n$, then

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + \sum_{i=1}^{n}(\bar{x} - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - a)^2$$

Sample Variance

By the algebraic equality, we have

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

Because of I.I.D., we have

Unbiased estimator for σ^2

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

$$\begin{aligned}(n-1)E[S^2] = \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] &= \sum_{i=1}^{n}\left(\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2\right) + \sum_{i=1}^{n}\left(\mathbb{E}[X_i]^2 - E[\bar{X}^2]\right) \\ &= n\mathrm{Var}(X_1) + n\left(\mathbb{E}[\bar{X}]^2 - E[\bar{X}^2]\right) \\ &= n\sigma^2 - n\mathrm{Var}(\bar{X}) = (n-1)\sigma^2\end{aligned}$$

Statistic and Parameter

- We already see that how the statistics, $\bar{X}$ and S^2 (why are they statistics?), are related to, but do not directly depend on, the parameters μ and σ^2.
- Need to be careful: which is random/deterministic, known/unknown?

Sample Mean and Variance

- How to calculate the distribution of the sample mean and sample variance?

Sample Mean – Convolution

Convolution

Suppose X and Y are independent continuous random variables with PDFs f_X and f_Y. The PDF of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z - w) dw$$

Proof by conditioning or Jacobian for $(Z, W) = (X + Y, X)$.

Moment Generating Function

We have learned how to compute $E[g(X)]$ if we know the distribution of X.

Now, consider $g(x) = e^{tx}$. (Think of t as a parameter.)

Moment generating function

The moment generating function $\phi(t)$ of the random variable X is defined as:

$$\phi(t) = E[e^{tX}]$$

- MGF is not always finite.

The Derivative of MGF

Why is MGF useful?

- Taking the first derivative of $\phi(t)$,[1]

$$\phi'(t) = \frac{d}{dt}E[e^{tX}] = E[\frac{d}{dt}e^{tX}] = E[Xe^{tX}]$$

Plug in $t = 0$,

$$\phi'(0) = E[Xe^0] = E[X]$$

- Taking the second derivative of $\phi(t)$,

$$\phi''(t) = \frac{d}{dt}\phi'(t) = \frac{d}{dt}E[Xe^{tX}] = E[\frac{d}{dt}Xe^{tX}] = E[X^2e^{tX}]$$

Plug in $t = 0$,

$$\phi''(0) = E[X^2e^0] = E[X^2]$$

[1] assuming that the interchange of $\mathbb{E}$ and d/dt is justified. More on this later.

Moment Generating Function

In general, the nth derivative of $\phi(t)$, evaluated at 0 equals the nth moment:

Generating moments

$$\phi^{(n)}(0) = E[X^n]$$

This is why $\phi(t)$ is called the moment generating function.

- Even when $E[X^n]$ exist for all n, it is still possible that $\phi(t)$ is not finite for all $t > 0$.

Interchanging the sign

In the MGF analysis, we have used $\frac{d}{dt}\mathbb{E}[\cdot] = \mathbb{E}[\frac{d}{dt}\cdot]$. When is this allowed?

- $\mathbb{E}$ is an integral, $\frac{d}{dt}$ is a limit.
- More generally, when do we have

$$\lim_{y \to y_0} \int_{-\infty}^{\infty} h(x, y) dx = \int_{-\infty}^{\infty} \lim_{y \to y_0} h(x, y) dx?$$

- This is not always true.
  Example: Consider $h(x, n) = \frac{1}{2n} \mathbb{1}_{-n \leq x \leq n}$. Then $\int_{-\infty}^{\infty} h(x, n) dx = 1$ but $\int_{-\infty}^{\infty} h(x, \infty) dx = 0$.

Dominated Convergence Theorem

Theorem

Suppose $h(x,y)$ *is continuous at* y_0 *for each* x*, and there exists* $g(x)$ *such that*

- $|h(x,y)| \leq g(x)$ *for all* x *and* y*; and*
- $\int_{-\infty}^{\infty} g(x)dx < \infty$*, then*

$$\lim_{y \to y_0} \int_{-\infty}^{\infty} h(x,y)dx = \int_{-\infty}^{\infty} \lim_{y \to y_0} h(x,y)dx.$$

Corollary

Suppose $f(x,t)$ is differentiable in t and there exits a $g(x,t)$ such that

- $|\frac{\partial f}{\partial t}(x,t_0)| \le g(x)$ *for all x and $|t-t_0|<\delta$; and*
- $\int_{-\infty}^{\infty} g(x)dx < \infty$, *then*

$$\frac{d}{dt}\int_{-\infty}^{\infty} f(x,t)dx = \int_{-\infty}^{\infty} \frac{\partial f(x,t)}{\partial t}dx.$$

Back to our example of $\frac{d}{dt}\mathbb{E}[f(X,t)]$, if $|\frac{d}{dt}f(x,t)|$ is bounded above by $g(x)$ and $\mathbb{E}[g(X)]$ exists, then we can interchange the order.

Try to verify the condition for the MGFs given that $\phi(t)<\infty$ for $|t|<\delta$. Hint: consider $X>0$ first and let $g(x)=xe^{\delta x/2}f(x)$.

Identically Distributed

Theorem

Let $F_X(\cdot)$ and $F_Y(\cdot)$ be two CDFs all of whose moments exits.

a. *If X and Y have bounded support, then $F_X \equiv F_Y$ if and only if $\mathbb{E}[X^r] = \mathbb{E}[Y^r]$ for all $r = 0, 1, 2, \ldots$*

b. *If the MGF's exist and $\phi_X(t) = \phi_Y(t)$ for all t in some neighborhood of 0, then $F_X \equiv F_Y$.*

Read Example 2.3.10 in Casella and Berger for a case when $\mathbb{E}[X_1^r] = \mathbb{E}[X_2^r]$ for all r but $F_1 \neq F_2$.

Convergence in Distribution

Consider two random variables X, Y with CDF F_X, F_Y and MGF $\phi_X(\cdot), \phi_Y(\cdot)$.

If

$$\phi_X(\cdot) \approx \phi_Y(\cdot),$$

can we assert that

$$F_X \approx F_Y?$$

Binomial

$B(n, p)$: suppose we have n independent trials, each with the same success probability p.

Let X be the number of successes out of these n trial

$$\mathbb{P}\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, 2, \ldots, n$$

- Called binomial because of the binomial expansion $(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}$.

Binomial MGF

Example: Binomial MGF

$$\phi(t) = \sum_{i=0}^{n} e^{ti} \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=0}^{n} \binom{n}{i} (pe^t)^i (1-p)^{n-i} = [pe^t + (1-p)]^n$$

- If

$$X_1 \sim \text{Binomial}(n_1, p), \quad X_2 \sim \text{Binomial}(n_2, p)$$

and X_1 and X_2 are independent, then

$$\phi_{X_1+X_2}(t) = E[e^{tX_1} e^{tX_2}] = E[e^{tX_1}]E[e^{tX_2}] = \phi_{X_1}(t)\phi_{X_2}(t) = [pe^t + (1-p)]^{n_1+n_2}.$$

Hence,

$$X_1 + X_2 \sim \text{Binomial}(n_1 + n_2, p)$$

Poisson

A random variable X is Poisson with parameter $\lambda > 0$ if the PMF is

$$\mathbb{P}\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}, \quad i = 0, 1, 2, \ldots$$

Example: Poisson MGF

$$\phi(t) = E[e^{tX}] = \sum_{i=0}^{\infty} e^{ti}e^{-\lambda}\lambda^i/i! = e^{-\lambda}\sum_{i=0}^{\infty}(e^t)^i\lambda^i/i!$$
$$= e^{-\lambda}\sum_{i=0}^{\infty}(\lambda e^t)^i/i! = e^{-\lambda}e^{\lambda e^t}$$

Similar to Binomial distribution, we can check that Poisson(λ_1)+Poisson(λ_2) is Poisson$(\lambda_1 + \lambda_2)$.

Poisson Cont.

Differentiation yields

$$\phi'(t) = e^{-\lambda}\lambda e^t e^{\lambda e^t}$$
$$\phi''(t) = e^{-\lambda}[(\lambda e^t)^2 e^{\lambda e^t} + \lambda e^t e^{\lambda e^t}]$$

So

$$E[X] = \phi'(0) = e^{-\lambda}\lambda e^0 e^{\lambda e^0} = \lambda$$
$$E[X^2] = \phi''(0) = e^{-\lambda}[(\lambda e^0)^2 e^{\lambda e^0} + \lambda e^0 e^{\lambda e^0}] = \lambda^2 + \lambda$$

Thus,

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Poisson and Binomial

- Note that $\phi_B(t) = (pe^t + (1-p))^n$ and $\phi_P(t) = e^{\lambda(e^t - 1)}$.
- Let $p = \lambda/n$ and $n \to \infty$

$$\phi_B(t) = (pe^t + (1-p))^n = \left(1 + \frac{1}{n}(e^t - 1)\lambda\right)^n$$
$$\longrightarrow e^{\lambda(e^t-1)} = \phi_P(t).$$

Implications

Now consider why the following real-world events are appropriate to be modeled as Poisson R.V.s

- The number of misprints in a book.
- The number of people in a community living to 100 years of age.
- The number of transistors that fail on their first day of use.
- The number of customers entering a post office on a given day.
- ...

Convergence in Distribution

Theorem

Consider a sequence of random variables $X_1, X_2, \ldots,$ *with CDF* $F_1, F_2, \ldots$ *and MGF* $\phi_1(\cdot), \phi_2(\cdot), \ldots$. *If there exists* $\delta > 0$ *such that*

$$\lim_{i\to\infty} \phi_i(t) = \phi(t), \quad \text{for all } t \in (-\delta, \delta),$$

where $\phi(\cdot)$ *is the MGF for* X *with CDF* F, *then for all* $x \in \mathbb{R}$

$$\lim_{i\to\infty} F_i(x) = F(x).$$

Sample Mean – MGF

MGF of sample mean

Suppose the MGF of F is ϕ, then the MGF of sample mean $\bar{X}$ is

$$\phi_{\bar{X}}(t) = [\phi(t/n)]^n.$$

Standard Normal

Standard Normal, $\mathcal{N}(0,1)$, has the density function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty$$

and cumulative distribution function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy, \quad -\infty < x < \infty$$

Let $Z \sim \mathcal{N}(0,1)$, what is the distribution of $X = \mu + \sigma Z$?

$$P\{X < x\} = P\left\{\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}\right\} = P\left\{Z < \frac{x-\mu}{\sigma}\right\} = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$$f_X(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

Location and Scale Families

Location and Scale Families

If $f(x)$ is a PDF, then $g(x|\mu,\sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ is a PDF

- The location parameter μ and the scale parameter σ.
- If $Z \sim f(z)$, then $X = \sigma Z + \mu \sim g(x)$.

Normal

A random variable is said to be normally distributed with parameters μ and σ^2, and we write $X \sim \mathcal{N}(\mu, \sigma^2)$, if the PDF is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

$$\mathbb{E}[X] = \mu, \quad \mathbb{E}[X^2] = \sigma^2 + \mu^2, \quad \mathrm{Var}(X) = \sigma^2.$$

Normal

We want to verify

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 1$$

Note that

$$\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} e^{-\frac{y^2}{2}} dxdy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dxdy = \int_{0}^{2\pi} \int_{0}^{\infty} re^{-\frac{r^2}{2}} drd\theta = \int_{0}^{2\pi} \int_{0}^{\infty} e^{-u} dud\theta$$

$$= \int_{0}^{2\pi} \int_{0}^{\infty} e^{-u} dud\theta = 2\pi$$

Normal

Finding mean and variance by direct computing

$$
\begin{aligned}
&\mathbb{E}[X-\mu] \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(x-\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{x-\mu}{\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \\
&= \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty}ye^{-\frac{y^2}{2}}dy \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{Var}(X) \\
&= \mathbb{E}[(X-\mu)^2] \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(x-\mu)^2e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \\
&= \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{(x-\mu)^2}{\sigma^2}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \\
&= \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}y^2e^{-\frac{y^2}{2}}dy \quad \text{(int by part)} \\
&= \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{y^2}{2}}dy \\
&= \sigma^2
\end{aligned}
$$

Normal MGF

For normal distribution, its MGF is

$$\begin{aligned}\phi(t) &= \int_{-\infty}^{\infty} \frac{e^{tx}}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2 - 2t\sigma^2 x}{2\sigma^2}} dx \\ &= \int_{-\infty}^{\infty} \frac{e^{\mu t + \sigma^2 t^2/2}}{\sqrt{2\pi}\sigma} e^{-\frac{(x-(\mu+t\sigma^2))^2}{2\sigma^2}} dx = e^{\mu t + \frac{\sigma^2 t^2}{2}}.\end{aligned}$$

Its moments are easy to compute

- $\phi'(t) = (\mu + \sigma^2 t)\phi(t)$, hence $E[X] = \mu$.
- $\phi''(t) = \sigma^2 \phi(t) + (\mu + \sigma^2 t)\phi'(t)$, hence $E[X^2] = \sigma^2 + \mu^2$ and $\mathrm{Var}(X) = \sigma^2$.
- $E[X^3]$ and $E[X^4]$ can be computed as well.

Sum of Independent Normal R.V.s

Suppose $X_1, X_2, \cdots, X_n$ are independent and $\mathcal{N}(\mu_i, \sigma_i^2)$.
The sum is still normal

$$X = \sum_i X_i \sim \mathcal{N}(\mu, \sigma^2)$$

where

$$\mu = \sum_i \mu_i, \quad \text{and} \quad \sigma^2 = \sum_i \sigma_i^2$$

The MGF of $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ is $\mathbb{E}[e^{tX_i}] = e^{\mu_i t + \frac{\sigma_i^2 t^2}{2}}$.

The MGF of X is

$$\begin{aligned} \mathbb{E}[e^{t\sum_{i=1}^n X_i}] &= \mathbb{E}[e^{tX_1} e^{tX_2} \cdots e^{tX_n}] = \mathbb{E}[e^{tX_1}]\mathbb{E}[e^{tX_2}] \cdots \mathbb{E}[e^{tX_n}] \qquad \text{by independence} \\ &= e^{\mu_1 t + \frac{\sigma_1^2 t^2}{2} + \mu_2 t + \frac{\sigma_2^2 t^2}{2} + \cdots \mu_n t + \frac{\sigma_n^2 t^2}{2}} = e^{\mu t + \frac{\sigma^2 t^2}{2}}. \end{aligned}$$

Multivariate Normal Distribution

Let $\mathbf{X} = (X_1, \ldots, X_k)$ be a random vector and let $\mathbf{\Sigma}$ be its covariance matrix, i.e., $\mathbf{\Sigma}_{i,j} = \mathrm{Cov}(X_i, X_j)$. Joint PDF $\mathbf{x} = (x_1, \ldots, x_k)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{k}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)$$

Bivariate: $\mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$

$$f_X(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right)$$

Properties of Multivariate Normal

- MGF $\mathbb{E}[e^{\mathbf{t}\cdot\mathbf{X}}] = \exp(\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t})$.
- When $\boldsymbol{\Sigma}$ is not full rank, then the distribution is degenerate (linearly dependent, not an interested case). Think of $(X, -X)$.
- When $\boldsymbol{\Sigma}_{ij} = 0$, then X_i and X_j are independent. For *joint* normal distributions, $\mathrm{Cov} = 0$ implies independence.
- The rule above only applies when the joint distribution is normal. Two marginal normal R.V.s may not be joint normal.

  Example: Let $W = 1$ w.p. 0.5 and $W = -1$ w.p. 0.5. Consider $Y = WX$.
- Linear transformation $A\mathbf{X}$ where $A \in \mathbb{R}^{m\times k}$. Then it is jointly normal with mean $A\boldsymbol{\mu}$ and variance matrix $A\boldsymbol{\Sigma}A^T$.
- The conditional distribution of $(X_1, \ldots, X_i)$ given $(X_{i+1}, \ldots, X_k)$ is still normal and has a closed form (check the wiki page).

Normal Sample Mean and Variance

Theorem (Normal Sample)

If $F \sim \mathcal{N}(\mu, \sigma^2)$, then

1. $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$
2. $\bar{X}$ *and* S *are independent*
3. $(n-1)S^2/\sigma^2$ *is* χ^2_{n-1}

Part one is straightforward. Either by MGF, or by the linear transformation of multivariate normal RVs.

Independence of Sample Mean and Sample Variance

- For part two, consider standard normal. Suffices to show $\bar{X}$ is independent of $(X_2 - \bar{X}, X_3 - \bar{X}, \ldots, X_n - \bar{X})$.
- Why $X_1 - \bar{X}$ is not needed? Because $\sum_{i=1}^n (X_i - \bar{X}) = 0$.
- $(\bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$ is a linear transformation $A\mathbf{X}$, where

$$A = \begin{pmatrix} 1/n & 1/n & \cdots & 1/n \\ -1/n & 1-1/n & \ldots & -1/n \\ \vdots & \vdots & \ddots & \vdots \\ -1/n & -1/n & \cdots & 1-1/n \end{pmatrix}$$

- In general, let $\boldsymbol{X} \sim \mathcal{N}(\mu, \Sigma)$, consider a matrix $A \in \mathbb{R}^{m \times n}$ and $A\mathbf{X}$ then $A\boldsymbol{X}$ is jointly normal with mean $A\mu$ and covariance matrix $A\Sigma A^T$.
- Need to show AA^T has zero entries on the first row/column (except for the diagonal).

Independence of Sample Mean and Sample Variance

- Alternatively, consider the transformation $y_1 = \bar{x}$, $y_i = x_i - \bar{x}$, $i = 2, \ldots, n$, with Jacobian $1/n$.
- Because $f_X(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2}$, we have

$$
\begin{aligned}
f_Y(y_1, \ldots, y_n) &= \frac{n}{(2\pi)^{n/2}} e^{-\frac{1}{2}(y_1 - \sum_{i=2}^n y_i)^2} e^{-\frac{1}{2}\sum_{i=2}^n (y_i + y_1)^2} \\
&= \left[\left(\frac{n}{2\pi} \right)^{1/2} e^{-\frac{1}{2} n y_1^2} \right] \left[\frac{n^{1/2}}{(2\pi)^{(n-1)/2}} e^{-\frac{1}{2}[\sum_{i=2}^n y_i^2 + (\sum_{i=2}^n y_i)^2]} \right]
\end{aligned}
$$

Chi-squared Distribution

For part three, let's first define Chi-squared distributions.

Definition

If $X_1, \ldots, X_k$ are independent standard normal, then $Q = \sum_{i=1}^{k} X_i^2 \sim \chi_k^2$ has chi-squared distribution with k degrees of freedom.

- Mean: k
- PDF:

$$f(x|k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

- Additivity: $\chi_{k_1}^2 + \chi_{k_2}^2 = \chi_{k_1+k_2}^2$.

$(n-1)S^2/\sigma^2$ Is Chi-Squared

Lemma

Let A be a symmetric matrix such that $A^2 = A$, and let $r = \text{trace}(A)$ denote the sum of the eigenvalue of A. If $X \sim N(0, \sigma^2 I)$ is a standard normal random vector, then

$$\frac{X^T A X}{\sigma^2} \sim \chi^2_r.$$

$$(n-1)S^2/\sigma^2 = \left(\frac{X_1 - \bar{X}}{\sigma}, \ldots, \frac{X_n - \bar{X}}{\sigma}\right)^T \left(\frac{X_1 - \bar{X}}{\sigma}, \ldots, \frac{X_n - \bar{X}}{\sigma}\right) = Z^T A^T A Z,$$

where

$$A = \begin{pmatrix} 1 - 1/n & -1/n & \cdots & -1/n \\ -1/n & 1 - 1/n & \ldots & -1/n \\ \vdots & \vdots & \ddots & \vdots \\ -1/n & -1/n & \cdots & 1 - 1/n \end{pmatrix}$$

Alternative Proof

- $(n-1)S^2$ is a sum of n linearly dependent normal RV squared. Want to show $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Set $\sigma = 1$ for simplicity.
- Use induction
  - The basis: When $n = 2$, $S_2^2 = (X_2 - X_1)^2/2$ is χ_1^2 (why?)
  - Inductive step:

$$(n-1)S_n^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2 = \left(\sum_{i=1}^{n-1} X_i^2 - (n-1)\bar{X}_{n-1}^2\right) + \frac{n-1}{n}(X_n - \bar{X}_{n-1})^2.$$

(1) $\frac{n-1}{n}(X_n - \bar{X}_{n-1})^2$ is a standard normal R.V. squared,
(2) $(X_n, \bar{X}_{n-1})$ is independent of S_{n-1}.

Can also be verified by evaluating the MGF of S_n^2 and χ^2_{n-1}.

Implication

- Although both $\bar{X}$ and $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$ have $\bar{X}$ in it, they are independent.
- Sample variance indeed measures the "spread" without affected by the "center".
- $(n-1)S^2$ has n squared normal RVs. They are correlated, have mean zero and variance $1 - 1/n$. Nevertheless, the sum is χ^2_{n-1}!

Student's t

It is easy to see that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is standard normal, what about

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} = \frac{U}{\sqrt{V/(n-1)}}$$

This is called a Student's t distribution with degree of freedom $n-1$.

- How to obtain its PDF? Joint PDF of U and V (independence), Jacobian, then marginal. Or condition on one PDF and then integrate.
- t_p has only up to $(p-1)$th moment.
- When p is large ($p \geq 20$), t_p is pretty much the same as a standard normal.

Snedecor's F

If $(X_1, \ldots, X_n)$ from $N(\mu_X, \sigma_X^2)$ and $(Y_1, \ldots, Y_m)$ from $N(\mu_Y, \sigma_Y^2)$, want to compare σ_X^2/σ_Y^2, but can only observe S_X^2/S_Y^2.

F distribution

$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$ has F distribution with $n-1$ and $m-1$ degrees of freedom. (the ratio of two χ^2)

- An $F(p,q)$ distribution is essentially $\frac{\chi_p^2/p}{\chi_q^2/q}$.
- If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$.
- If $X \sim t_q$, then $X^2 \sim F_{1,q}$.

Order Statistics

Definition

Order statistics $(X_{(1)}, \dots, X_{(n)})$ is the ascending order of random sample $(X_1, \dots, X_n)$.

- Sample range $X_{(n)} - X_{(1)}$.
- Sample median
  - $X_{((n+1)/2)}$ if n is odd
  - $\left(X_{(n/2)} + X_{(n/2+1)}\right)/2$ if n is even
- The $(100p)^{th}$ percentile is $X_{(\lfloor np \rfloor)}$

Order Statistics

Theorem (Discrete)

Suppose the PMF is $f(x_i) = p_i$ for $x_1 < x_2 < \dots$. Define $P_0 = 0$ and $P_i = \sum_{j=1}^{i} p_j$. Then,

$$\mathbb{P}(X_{(j)} \le x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}$$

$$\mathbb{P}(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k} \right]$$

$$\mathbb{P}(X_{(j)} \le x_i) = \mathbb{P}(\text{at least } j \text{ samples are less than equal to } x_i).$$

Order Statistics

Theorem (Continuous)

Suppose the PDF is f and cdf is F. Then,

$$F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k}$$

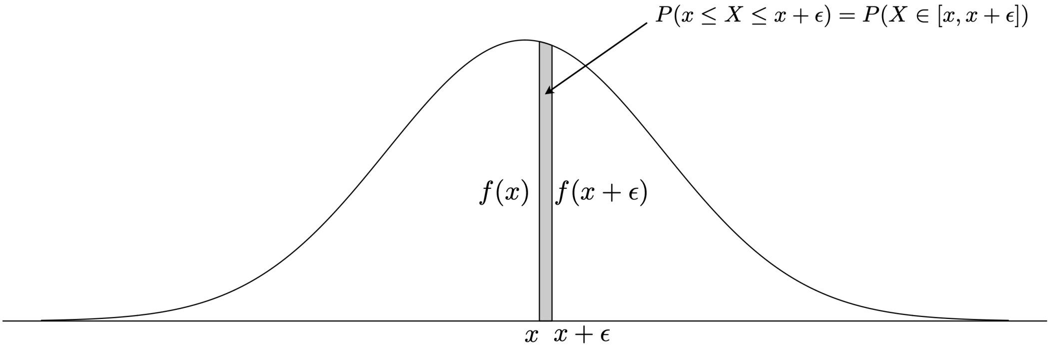
$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f(x) [F(x)]^{j-1} [1 - F(x)]^{n-j}$$

For density, take derivative and use

$$\frac{d}{dp} \sum_{k=j}^{n} \binom{n}{k} p^k (1-p)^{n-k} = n \binom{n-1}{j-1} p^{j-1} (1-p)^{n-k}.$$

Order Statistics

Heuristic



Order Statistics

For Uniform $(0,1)$,

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!}x^{j-1}(1-x)^{n-j}$$
$$= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)}x^{j-1}(1-x)^{(n-j+1)-1} \sim \mathrm{Beta}(j, n-j+1)$$

Order Statistics

Theorem (Continuous – Joint)

Suppose the PDF is f and cdf is F. Then the joint PDF,

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f(u)f(v)$$
$$[F(u)]^{i-1}[F(v)-F(u)]^{j-1-i}[1-F(v)]^{n-j}, \quad u<v$$

- Informal proof:
  $f_{i,j}(u,v) = \mathbb{P}(i-1$ less than u, $n-j$ greater than v, one at u, one at $v)$.
- Be careful about the domain!
- $f_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots,x_n) = n! f(x_1)\cdots f(x_n)$, $x_1 < x_2 < \cdots < x_n$.

Order Statistics

Example: Consdier range: $R = X_{(n)} - X_{(1)}$, and midrange: $V = (X_{(n)} + X_{(1)})/2$

For Uniform $(0, a)$, the joint PDF

$$f_{X_{(1)},X_{(n)}}(x_1, x_n) = \frac{n(n-1)}{a^2}\left(\frac{x_n}{a} - \frac{x_1}{a}\right)^{n-2} = \frac{n(n-1)(x_n - x_1)^{n-2}}{a^n}, \quad 0 < x_1 < x_n < a.$$

$X_{(1)} = V - R/2$, $X_{(n)} = V + R/2$

The joint distribution for (R, V) is

$$f_{R,V}(r, v) = \frac{n(n-1)r^{n-2}}{a^n}, \quad 0 < r < a,\ \frac{r}{2} < v < a - \frac{r}{2}.$$

Convergence Modes

In general, real-valued random variables (can be extended to $\mathbb{R}^n$) $X_1, X_2, \ldots$ can converge to another random variable X in several different modes:

Definition (Convergence in Distribution)

$$X_n \Rightarrow X: \lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad \text{for all } x \text{ where } F_X \text{ is continuous.}$$

Definition (Convergence in Probability)

$$X_n \xrightarrow{p} X: \lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0, \quad \text{for any } \epsilon > 0.$$

Definition (Convergence in Almost Surely)

$$X_n \xrightarrow{a.s.} X: \mathbb{P}(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1, \quad \text{for any } \epsilon > 0.$$

Connection

- a.s. convergence is stronger than convergence in probability, which is stronger than convergence in distribution.
- Can you find examples of convergence in P but not a.s.? convergence in distribution but not in P?

Continuous mapping theorem

A continuous mapping ($g : \mathbb{R} \to \mathbb{R}$) preserves convergence in all the three modes.

in probability v.s. almost surely

Construct a sample space $S = [0,1]$ equipped with the Borel σ-algebra and uniform distribution. Construct random variables

$$X_1(s) = 1_{\{s\in[0,1]\}},\ X_2(s) = 1_{\{s\in[0,\frac{1}{2}]\}},\ X_3(s) = 1_{\{s\in[\frac{1}{2},1]\}},$$
$$X_4(s) = 1_{\{s\in[0,\frac{1}{3}]\}},\ X_5(s) = 1_{\{s\in[\frac{1}{3},\frac{2}{3}]\}},\ X_6(s) = 1_{\{s\in[\frac{2}{3},1]\}},$$

Convergence in probability? Yes! Convergence almost surely? No!

in probability v.s. in distribution

X is standard normal. $X_n = -X$.

Theorem

$\xrightarrow{p}$ and $\Rightarrow$ are equivalent when the limit is deterministic.

Let $X_{(n)}$ be the max of n independent Uniform $(0, 1)$.

$$\mathbb{P}(|X_{(n)} - 1| \geq \epsilon) = \mathbb{P}(X_{(n)} \leq 1 - \epsilon) = (1 - \epsilon)^n \to 0$$

So we have convergence in probability. Let $\epsilon = t/n$

$$\mathbb{P}(n(1 - X_{(n)}) \leq t) = \mathbb{P}(X_{(n)} \leq 1 - t/n) = (1 - t/n)^n \to e^{-t}$$

So $n(1 - X_{(n)}) \Rightarrow$Exponential(1).

WLLN

Weak Law of Large Numbers

If $X_1, X_2, X_3, \ldots$ are i.i.d. with finite mean μ and variance σ^2, then

$$P\left\{\left|\frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} - \mu\right| > \epsilon\right\} \to 0 \quad \text{as } n \to \infty$$

$$E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \mu \quad \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

By Chebyshev's Inequality,

$$P\left\{\left|\frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2/n}{\epsilon^2} \to 0 \quad \text{as } n \to \infty$$

Consistent Estimators

Consistency

An estimator T_n is a consistent estimator for θ if T_n converges to θ in probability.

- When there are more samples, the statistic is becoming more "accurate".
- WLLN implies that the sample mean is consistent for μ.
- Consider the sample variance

$$S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}, \quad \mathbb{P}(|S_n^2 - \sigma^2| > \epsilon) \leq \frac{\mathbb{E}(S_n^2 - \sigma^2)^2}{\epsilon^2}$$

  So if $\mathrm{Var}(S_n^2) \to 0$ as $n \to \infty$ (true for normal), then the estimator is consistent.
- By continuous mapping ($g(x) = \sqrt{x}$), S_n is also a consistent estimator for σ (but biased by Jensen's inequality).

SLLN

Strong law of large numbers

If $X_1, X_2, X_3, \ldots$ are i.i.d. with finite mean μ and variance σ^2, then

$$\mathbb{P}\left(\lim_{n\to\infty} \left|\bar{X}_n - \mu\right| < \epsilon\right) = 1$$

- Both laws actually only require finite expected value. But the proof requires advanced measure theory.
- WLLN holds under even weaker conditions.

The Central Limit Theorem

Central Limit Theorem (CLT)

If $X_1, X_2, X_3, \ldots$ are i.i.d. with finite mean μ and variance σ^2, then

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \Rightarrow N(0,1).$$

Let $Y_i = (X_i - \mu)/\sigma$, the LHS is $\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$.

$$\phi_{LHS}(t) = \left(\phi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n = \left(1 + \frac{(t/\sqrt{n})^2}{2} + o(n^{-1})\right)^n \to e^{t^2/2}$$

The proof is based on the existence of MGF, but we can use characteristic functions when MGF does not exist.

Normal Approximates Binomial

Let $X \sim$Binomial(n, p), what happens if we fix p and let n grow large?

We can think of $X = \sum_{i=1}^{n} X_i$ with $X_i \sim$Bernoulli(p) for all i.

We want to approximate

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}\left(\frac{a - np}{\sqrt{np(1-p)}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{b - np}{\sqrt{np(1-p)}}\right)$$
$$= \Phi(\frac{b - np}{\sqrt{np(1-p)}}) - \Phi(\frac{a - np}{\sqrt{np(1-p)}})$$

Continuity correction

$$\mathbb{P}(X = a) = \mathbb{P}\left(\frac{a + \frac{1}{2} - np}{\sqrt{np(1-p)}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

Joint Convergence

Theorem (Slutsky's)

If $X_n \Rightarrow X$ and $Y_n \Rightarrow a$ (equivalently $Y_n \xrightarrow{p} a$), then

- $X_n Y_n \Rightarrow aX$
- $X_n + Y_n \Rightarrow X + a$

Implication:

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sigma}{S_n}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \Rightarrow N(0,1)$$

t_{n-1} distribution approximates standard normal.

Example: $X_n \sim \text{Uniform}(0,1)$ and $Y_n = -X_n$. The sum $X_n + Y_n = 0$ for all values of n. Moreover, $Y_n \Rightarrow \text{Uniform}(-1,0)$, but $X_n + Y_n$ does not converge in distribution to $X + Y$.

Delta Method

First-order Delta Method

Let Y_n be such that $\sqrt{n}(Y_n - \theta) \Rightarrow \mathcal{N}(0, \sigma^2)$. Suppose $g'(\theta)$ exists and is non-zero. Then

$$\sqrt{n}\,[g(Y_n) - g(\theta)] \Rightarrow \mathcal{N}(0, \sigma^2[g'(\theta)]^2).$$

Proof: Taylor expansion

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + o(Y_n - \theta)$$

Then use Slutsky's theorem.

Application: Estimating the odds

For Bernoulli(p) sample, we can use $\hat{p}_n = \sum_{i=1}^n X_i/n$ to estimate p. What about the odd $\frac{p}{1-p}$? Let $g(p) = \frac{p}{1-p}$, then $g'(p) = \frac{1}{(1-p)^2}$.

$$\mathbb{E}\left[\frac{\hat{p}_n}{1-\hat{p}_n}\right] \approx g(p) + g'(p)\mathbb{E}(\hat{p}_n - p) + \dots$$

$$\begin{aligned}\mathrm{Var}\left(\frac{\hat{p}_n}{1-\hat{p}_n}\right) &\approx \mathbb{E}[g(\hat{p}_n) - g(p)]^2 \approx \mathbb{E}[g'(p)(\hat{p}_n - p)]^2 \\ &\approx [g'(p)]^2 \mathrm{Var}(\hat{p}_n) \approx \frac{1}{(1-p)^4}\frac{p(1-p)}{n}\end{aligned}$$

Therefore,

$$\sqrt{n}\left(\frac{\hat{p}_n}{1-\hat{p}_n} - \frac{p}{1-p}\right) \Rightarrow \mathcal{N}(0, \frac{p}{(1-p)^3})$$

Delta Method

Second-order Delta Method

Let Y_n be such that $\sqrt{n}(Y_n - \theta) \Rightarrow \mathcal{N}(0, \sigma^2)$. Suppose $g'(\theta) = 0$ and $g''(\theta)$ exists and is non-zero. Then

$$n\left[g(Y_n) - g(\theta)\right] \Rightarrow \sigma^2 \frac{g''(\theta)}{2} \chi_1^2.$$