IEDA 2540 – Topic III: Properties of a Random Sample

Wei YOU

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

Spring, 2023

Population and Sample

Recall that

- A population is the subjects of interest of a study.
- A sample is a subset of the target population collected for statistical analysis.
- We have seem several ways to collect sample:
  - Simple random sampling, systematic sampling, cluster sampling, stratified sampling...

In today's lecture, we will focus on a specific sampling method that is most frequently used in statistics: random sample.

Infinite Population

Simple random sampling

- Deal with finite population, usually without replacement.
- Each element in the population has equal probability of being selected.

If the population is really large or even infinite

- E.g., the height of all people in the world.
- As an abstraction and simplification, we can treat it as a distribution F.
  - E.g., $F(x)$ is the fraction of people whose height is less than x.
- In other words, we regard the outcome of a random experiment as a random variable following the distribution F.

Almost all of the statistical analysis we will see in this class assumes that we observe sample from a distributions and infer the form of it.

Random Sample

Random Sample

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

[a] also called independent and identically distributed, i.i.d.

From probability theory, the i.i.d. property tell us that the joint PMF/PDF is

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

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

Example: If the population is finite, and the distributions is uniform, then random sample becomes the simple random sample (with replacement).

Sample from a Finite Population without Replacement

In reality, we usually sample without replacement. Is it still reasonable to assume i.i.d. sample?

- Without replacement, for finite population, the distributions of X_1 and X_2 are not exactly i.i.d.
  Example: There are 10 balls, numbering 1 to 10; we sample 2 from them.
- What if the population size is large?
  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).

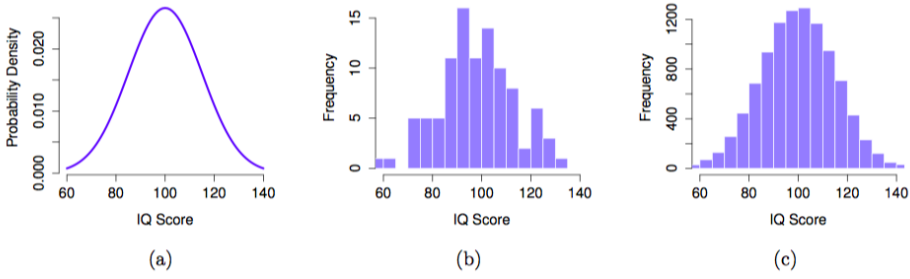
Sample from a Finite Population without Replacement

When the population is not too small, sampling with or without replacement are almost identical.

We focus on the i.i.d. random sample throughout the course unless stated otherwise.

Population Parameters and Sample Statistics

Consider the population distribution of IQ scores and two samples.



- We don't (but want to) know (a), at least its population parameters, such as mean and variance. If the sample is (c), then we have a good idea of (a).
- But in reality, we usually have (b), which does not look like (a). Can we find certain sample statistics of (b) to tell us about (a)?

Sample Statistics

Statistic – mathematical defintion

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! The parameters of interest are assumed to be unknown, otherwise we wouldn't need statistical analysis.

In this lecture, we will focus on three statistics and study their properties.

- Sample mean.
- Sample variance.
- Ordered statistics.

Sample Statistics

Suppose we have a sample from a population with distribution F

$$X_1, X_2, \ldots, X_n$$

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.

Sample Statistics

	Value known?	Value random?	Example
Parameter	unknow	deterministic	Population mean μ, variance σ^2
Statistic	known	random	Sample mean $\bar{X}$, sample variance S^2

The sample mean $\bar{X}$ is what we call an estimator of the parameter of interest μ, the population mean.

Are they useful?

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

Estimator

Estimator

An estimator is a rule for calculating an estimate of a quantity of interest based on observed data.

The following three are distinguished

- The quantity of interest: some parameter of the population.
- The estimator: a rule of calculation. Estimator is a statistic.
- The estimate: the value of the estimator, which depends on a given set of observation.

Properties of the Sample Mean – the Center

Unbiasedness

An estimator is said to be unbiased, if the expected value of the estimator is equal to the parameter it intends to estimate.

Sample mean is unbiased for the population mean μ

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

Properties of the Sample Mean – the Center

Mean squared error (MSE) measures the average squared-distance between your estimator and the sample observations.

Sample mean minimizes mean squared error (MSE)

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

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

A useful algebraic identity

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

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

Properties of the Sample Mean – the Variability

The center itself does not tell us about the dispersion/variability of our estimator.

Example: Consider two factors in the accuracy of measuring length: (a) the precision of the ruler; and (b) the number of measurements you take.

If F has finite variance σ^2, then

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

Properties of the Sample Mean – the Variability

As the sample size n increases, the sample mean becomes less and less random. Together with the fact that it is unbiased, we see that the sample mean converges to the population mean.

- This property is called consistency, which we shall see in a few slides.

Sample Mean – The Distritbution

Having seen the mean and variance of the sample mean, what about its distribution?

Convolution

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

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

The exact distribution of the sample mean (for small n) is usually hard to compute.

Exceptions: normal (we will see later), Poisson and other exponential families.

Consistency Fail to Hold – Cauchy

Example: Consistency does not always hold. Recall that the sample mean satisfies

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \to 0, \qquad \text{as } n \to \infty.$$

One would expect that as long as we collect enough data, we can always get a precise estimation. Is it always true?

Cauchy distribution

A random variable is said to be a $\text{Cauchy}(0, \xi)$ random variable, if its pdf is

$$f(x) = \frac{1}{\pi\xi} \frac{1}{1 + (x/\xi)^2}.$$

- For i.i.d. $\text{Cauchy}(0, 1)$ Z_i, $\bar{Z}$ is still Cauchy$(0, 1)$!
- So we no longer observe the convergence as $n \to \infty$. Why is that?
- Because Cauchy random variables do not have finite means.

Properties of the Sample Variance – the Center

Now we look at the sample variance S^2, as an estimator for σ^2.

Question 1: is S^2 unbiased for σ^2?

An equivalent formula for sample variance

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

Note how it resembles

$$\sigma^2 = \mathsf{Var}(X) = E[(X-\mu)^2] = E[X^2] - \mu^2.$$

Proof. Expanding the square.

Properties of the Sample Variance – the Center

Is S^2 unbiased for σ^2?

$$(n-1)E[S^2] = \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] = \sum_{i=1}^{n}\left(\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2\right) + \sum_{i=1}^{n}\left(\mathbb{E}[X_i]^2 - E[\bar{X}^2]\right)$$

$$= n\mathrm{Var}(X_1) + \sum_{i=1}^{n}\left(\mathbb{E}[\bar{X}]^2 - E[\bar{X}^2]\right)$$

$$= n\sigma^2 - n\mathrm{Var}(\bar{X}) = (n-1)\sigma^2$$

Unbiased estimator for σ^2

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

Statistic and Parameters

We now see that how the statistics, $\bar{X}$ and S^2, are related to the parameters μ and σ^2.

- Need to be careful: which is random/deterministic, known/unknown?

	Value known?	Value random?	Example
Parameter	unknow	deterministic	Population mean μ, variance σ^2
Statistic	known	random	Sample mean $\bar{X}$, sample variance S^2

The Law of Large Numbers

Sometimes when we may think that the the statistics of a data set is still very irregular.

- One would say collect more data!
- Does this intuition make sense? Why?
- The law of large numbers ensures that large samples generally give you better information.

The Weak and Strong Laws of Large Numbers

Let's add a subscript n to denote the sample size (number of observations).

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

$$\mathbb{P}\left\{\left|\bar{X}_n - \mu\right| > \epsilon\right\} \to 0 \quad \text{as } n \to \infty.$$

*Converge in probability.

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

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

*Converge almost surely.

Consistent Estimators

From LLN, the sample mean converges to the population mean.

Consider the sample variance

$$
\begin{aligned}
S_n^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \\
&= \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right) \\
&= \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{n}{n-1}\bar{X}^2 \\
&\Rightarrow 1 \times \mathbb{E}[X^2] - 1 \times \mu^2 = \sigma^2
\end{aligned}
$$

By LLN, if $\mathrm{Var}(X^2)$ is finite, then S_n^2 converges to σ^2 as the sample size grows.

Consistent Estimators

We have seen that both the sample mean and sample variance converges to the parameter they estimates.

Consistency

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

- Consistency implies that having larger sample size helps.
- This is a desired property we want for most estimators.

S_n is also a consistent estimator for σ (*by continuous mapping theorem).

Beyond Sample Mean/variance

Example: Consider an unfair die provided by a casino. You want to check the probability of 6.

- Intuitively, you can roll it 1000 times, and count the number of 6s, and divide it by 1000.
- Does the intuition lead to accurate estimate?
- LLN can help. But how?

Empirical probability

The empirical probability of an event is the frequency of this event normalized by the total number of events.

Empirical Probability

Key observation

For an event A, consider the *random variable* $\mathbb{1}_{X_i \in A}$, then

$$\mathbb{P}(X_i \in A) = \mathbb{E}[\mathbb{1}_{X_i \in A}].$$

- The variance of $\mathbb{1}_{X_i \in A}$ is always finite. So LLN implies that

Convergence of the empirical probability

As $n \to \infty$, $\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \in A} \to \mathbb{E}[\mathbb{1}_{X_i \in A}] = \mathbb{P}(A)$ almost surely.

- To estimate the probability of an event, compute its empirical probability in a large sample.

Empirical Cumulative Distribution Function

Empirical probability works for continuous RV as well.

Example: What is the probability that a snake is shorter than 1 meter? Use $\mathbb{1}_{X_i<1}$.

We can conveniently combine these empirical probabilities into empirical distributions.

Empirical cumulative distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}.$$

- Empirical CDF is always discrete, even if X is continuous.
- By LLN, $\lim_{n\to\infty} F_n(x) = F(x)$ almost surely.

The Central Limit Theorem

Consider a normalized sample $Y_i = (X_i - \mu)/\sigma$. By LLN,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \Rightarrow 0.$$

However, if we use a different scale

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

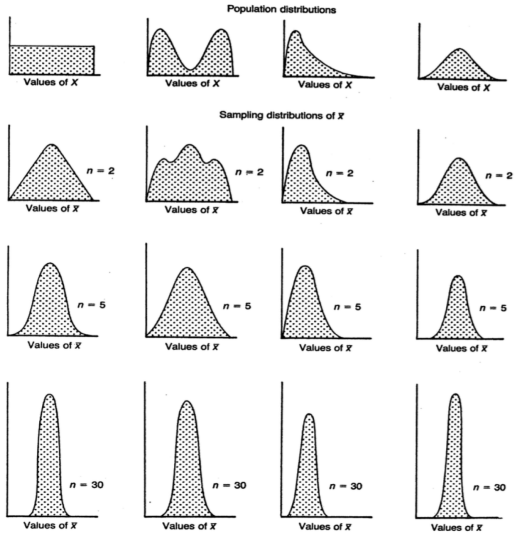
$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \Rightarrow N(0,1).$$

Normal Approximation

If the sample size n is "large", then the sum $\sum_{i=1}^{n} X_i$ is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

- The CLT can be applied regardless of the distribution of the individual values.
- The quality of the normal approximation varies.
  - If the underlying distribution is normal, then the approximation is perfect.
  - If the underlying distribution is skewed, then the approximation may be poor for smaller sample sizes.
- In reality, for distributions that are not too skewed, the rule of thumb is that $n = 30$ will usually suffice.

Normal Approximation



Example

Let $Y = \sum_{i=1}^{n} X_i$ denote the sum of the variables in a random sample of size $n = 30$ from the uniform distribution on $[0, 1]$.

Find normal approximations of $\mathbb{P}(13 < Y < 18)$.

- By CLT,

$$\frac{1}{\sqrt{30 \times (1/12)}}(Y - 30 \times 0.5) = \frac{1}{\sqrt{30}} \frac{\sum_{i=1}^{n}(X_i - 0.5)}{\sqrt{1/12}} \Rightarrow N(0, 1).$$

  So

$$\mathbb{P}(13 < Y < 18) \approx \mathbb{P}(-1.26 < Z < 1.90) \approx 0.87.$$

- The 90% percentile of Y.
  - First compute the 90% percentile of a standard normal RV Z, 1.28. Then transform it to get it for Y: $1.28 \times \sqrt{30/12} + 0.5 \times 30 \approx 17.0$.

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

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

Distribution of the Statistics

CLT states that normal distribution can be used to approximate sum or sample mean of random variables. What can we say about the distribution of the statistics for normal random sample?

Question: Why do we need the distribution of the statistics?

Example: Let's say that we suspect that a population have mean 0. How do we check if we are correct?

- We know that $\bar{X}$ is a good approximation of the population mean.
- We calculate $\bar{X}$ and compare it with our hypothesized value 0.
- If $\bar{X}$ is faraway from 0, it is not likely that $\mu = 0$.
- But how far is far enough?

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}

Chi-squared Distribution

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

Definition

If $X_1,\ldots,X_n$ 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: $\chi_{k_1}^2 + \chi_{k_2}^2 = \chi_{k_1+k_2}^2$.

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

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

Student's t distribution

Let U be a standard normal random variable and let V be a chi-squared random variable with degree of freedom $n-1$. Furthermore, assume that U and V are independent. Then $t_{n-1} = \frac{U}{\sqrt{V/(n-1)}}$ has a Student's t distribution with degree of freedom $n-1$.

- When n is large ($n \geq 20$), t_{n-1} 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

$F_{n-1,m-1} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$ has F distribution with $n-1$ and $m-1$ degrees of freedom.

- If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$.
- If $X \sim t_q$, then $X^2 \sim F_{1,q}$.

Normal Sample Proof – Part One, Distribution of Sample Mean

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

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

This solves part one.

Normal Sample Proof – Part Two, Independence of Sample Mean and Sample Variance*

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 $\mathbf{x} = (x_1, \ldots, x_k)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$

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

Important Properties of Normal RVs*

Theorem

Let $X_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$, $j = 1, \ldots, n$, independent.

$$U = \sum_{j=1}^{n} a_j X_j, \quad V = \sum_{j=1}^{n} b_j X_j.$$

U and V are independent if and only if $Cov(U, V) = \sum_{j=1}^{n} a_j b_j \sigma_j^2 = 0$.

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 $A^T \Sigma A$.

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

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

Need to show $A^T A$ has zero entries on the first row/column (except the diagonal).

Independence of Sample Mean and Sample Variance*

Alternative derivation of independence*

- 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

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

Normal Sample Proof – Part Three, S^2 Is Chi-Squared*

$(n-1)S^2$ 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:

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

(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}.

Order Statistics

Definition

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

- 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,*

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

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

$$\mathbb{P}(X_{(j)} \leq x_i) = \mathbb{P}(\text{at least } j \text{ samples are less than equal to } x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}$$

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 Uniform $(0, 1)$,

$$\begin{aligned} f_{X_{(j)}}(x) &= \frac{n!}{(j-1)!(n-j)!} x^{j-1} (1-x)^{n-j} \\ &= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1} (1-x)^{(n-j+1)-1} \sim \text{Beta}(j, n-j+1) \end{aligned}$$

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

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