Stat 5102 Lecture Slides
Deck 1

Charles J. Geyer
School of Statistics
University of Minnesota

Empirical Distributions

The *empirical distribution* associated with a vector of numbers $\mathbf{x} = (x_1, \ldots, x_n)$ is the probability distribution with expectation operator

$$E_n\{g(X)\} = \frac{1}{n} \sum_{i=1}^{n} g(x_i)$$

This is the same distribution that arises in finite population sampling. Suppose we have a population of size n whose members have values x_1, ..., x_n of a particular measurement. The value of that measurement for a randomly drawn individual from this population has a probability distribution that is this empirical distribution.

The Mean of the Empirical Distribution

In the special case where $g(X) = X$, we get the mean of the empirical distribution

$$E_n(X) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

which is more commonly denoted $\bar{x}_n$.

Those who have had another statistics course will recognize this as the formula of the *population mean*, if x_1, ..., x_n is considered a finite population from which we sample, or as the formula of the *sample mean*, if x_1, ..., x_n is considered a sample from a specified population.

The Variance of the Empirical Distribution

The variance of any distribution is the expected squared deviation from the mean of that same distribution. The variance of the empirical distribution is

$$
\begin{aligned}
\operatorname{var}_n(X) &= E_n\Big\{[X - E_n(X)]^2\Big\} \\
&= E_n\Big\{[X - \overline{X}_n]^2\Big\} \\
&= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2
\end{aligned}
$$

The only oddity is the use of the notation $\bar{x}_n$ rather than μ for the mean.

The Variance of the Empirical Distribution (cont.)

As with any probability distribution we have

$$\text{var}_n(X) = E_n(X^2) - E_n(X)^2$$

or

$$\text{var}_n(X) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - \bar{x}_n^2$$

The Mean Square Error Formula

More generally, we know that for any real number a and any random variable X having mean μ

$$E\{(X - a)^2\} = \text{var}(X) + (\mu - a)^2$$

and we called the left-hand side mse(a), the "mean square error" of a as a prediction of X (5101 Slides 32 and 33, Deck 2).

The Mean Square Error Formula (cont.)

The same holds for the empirical distribution

$$E_n\{(X - a)^2\} = \operatorname{var}_n(X) + (\bar{x}_n - a)^2$$

Characterization of the Mean

The mean square error formula shows that for any random variable X the real number a that is the "best prediction" in the sense of minimizing the mean square error $\text{mse}(a)$ is $a = \mu$.

In short, the mean is the best prediction in the sense of minimizing mean square error (5101 Slide 35, Deck 2).

Characterization of the Mean (cont.)

The same applies to the empirical distribution. The real number a that minimizes

$$E_n\{(X - a)^2\} = \frac{1}{n} \sum_{i=1}^{n} (x_i - a)^2$$

is the mean of the empirical distribution $\bar{x}_n$.

Probability is a Special Case of Expectation

For any random variable X and any set A

$$\Pr(X \in A) = E\{I_A(X)\}$$

If P is the probability measure of the distribution of X, then

$$P(A) = E\{I_A(X)\}, \qquad \text{for any event } A$$

(5101 Slide 63, Deck 1).

Probability is a Special Case of Expectation (cont.)

The same applies to the empirical distribution. The probability measure P_n associated with the empirical distribution is defined by

$$\begin{aligned} P_n(A) &= E_n\{I_A(X)\} \\ &= \frac{1}{n}\sum_{i=1}^{n} I_A(x_i) \\ &= \frac{\text{card}\{\, i : x_i \in A \,\}}{n} \end{aligned}$$

where card(B) denotes the *cardinality* of the set B (the number of elements it contains).

Probability is a Special Case of Expectation (cont.)

In particular, for any real number x

$$P_n(\{x\}) = \frac{\text{card}\{\, i : x_i = x \,\}}{n}$$

One is tempted to say that the empirical distribution puts probability $1/n$ at each of the points x_1, ..., x_n, but this is correct only if the points x_1, ..., x_n are distinct.

The statement that is always correct is the one above. The empirical distribution defines the probability of the point x to be $1/n$ times the number of i such that $x_i = x$.

Empirical Distribution Function

In particular, the distribution function (DF) of the empirical distribution is defined by

$$
\begin{aligned}
F_n(x) &= P_n(X \leq x) \\
&= \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(x_i) \\
&= \frac{\text{card}\{\, i : x_i \leq x \,\}}{n}
\end{aligned}
$$

Order Statistics

If

$$x_1, x_2, \ldots, x_n$$

are any real numbers, then we use the notation

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)} \qquad (*)$$

for the same numbers put in sorted order so $x_{(1)}$ is the least and $x_{(n)}$ the greatest.

Parentheses around the subscripts denotes sorted order.

These numbers $(*)$ are called the *order statistics*.

Quantiles

If X is a random variable and $0 < q < 1$, then the q-*th quantile* of X (or of the distribution of X) is any number x such that

$$\Pr(X \leq x) \geq q \quad \text{and} \quad \Pr(X \geq x) \geq 1 - q$$

If X is a discrete random variable having distribution function F, then this simplifies to

$$F(y) \leq q \leq F(x), \qquad y < x$$

(5101 Slide 60, Deck 4).

Quantiles of the Empirical Distribution

The q-*th quantile* of the empirical distribution is any number x such that

$$P_n(X \leq x) \geq q \quad \text{and} \quad P_n(X \geq x) \geq 1 - q$$

or such that

$$F_n(y) \leq q \leq F_n(x), \qquad y < x$$

(the two conditions are equivalent).

Quantiles of the Empirical Distribution (cont.)

For any real number a the notation $\lceil a \rceil$ (read "ceiling of a") denotes the least integer greater than or equal to a. For any real number a the notation $\lfloor a \rfloor$ (read "floor of a") denotes the greatest integer less than or equal to a.

If nq is not an integer, then the q-th quantile is unique and is equal to $x_{(\lceil nq \rceil)}$

If nq is an integer, then the q-th quantile is not unique and is any real number x such that

$$x_{(nq)} \leq x \leq x_{(nq+1)}$$

Quantiles of the Empirical Distribution (cont.)

(nq not an integer case). Define $a = x_{(\lceil nq \rceil)}$, the number we are to show is the empirical q-th quantile. There are at least $\lceil nq \rceil$ of the x_i less than or equal to a, hence

$$P_n(X \leq a) \geq \frac{\lceil nq \rceil}{n} \geq q$$

There are at least $n - \lceil nq \rceil + 1$ of the x_i greater than or equal to a, hence

$$P_n(X \geq a) \geq \frac{n - \lceil nq \rceil + 1}{n} = \frac{n - \lfloor nq \rfloor}{n} \geq 1 - q$$

Quantiles of the Empirical Distribution (cont.)

(nq is an integer case). Define a to be any real number such that

$$x_{(nq)} \leq a \leq x_{(nq+1)}$$

We are to show that a is an empirical q-th quantile. There are at least nq of the x_i less than or equal to a, hence

$$P_n(X \leq a) \geq \frac{nq}{n} = q$$

There are at least $(n+1) - (nq+1) = n(1-q)$ of the x_i greater than or equal to a, hence

$$P_n(X \geq a) \geq \frac{n(1-q)}{n} = 1 - q$$

Quantiles of the Empirical Distribution (cont.)

Suppose the order statistics are

0.03	0.04	0.05	0.49	0.50	0.59	0.66	0.72	0.83	1.17

Then the 0.25-th quantile is $x_{(3)} = 0.05$ because $\lceil nq \rceil = \lceil 2.5 \rceil = 3$. And the 0.75-th quantile is $x_{(8)} = 0.72$ because $\lceil nq \rceil = \lceil 7.5 \rceil = 8$. And the 0.5-th quantile is any number between $x_{(5)} = 0.50$ and $x_{(6)} = 0.59$ because $nq = 5$ is an integer.

Empirical Median

Nonuniqueness of empirical quantiles can be annoying. People want one number they can agree on. But there is, for general q, no such agreement.

For the median (the 0.5-th quantile) there is widespread agreement. Pick the middle number of the interval.

If n is odd, then the *empirical median* is the number

$$\tilde{x}_n = x_{(\lceil n/2 \rceil)}$$

If n is even, then the *empirical median* is the number

$$\tilde{x}_n = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

Characterization of the Median

For any random variable X, the median of the distribution of X is the best prediction in the sense of minimizing mean absolute error (5101 Slides 70–74, Deck 4).

The median is any real number a that minimizes

$$E\{|X - a|\}$$

considered as a function of a.

Characterization of the Empirical Median

The empirical median minimizes

$$E_n\{|X - a|\} = \frac{1}{n} \sum_{i=1}^{n} |x_i - a|$$

considered as a function of a.

Characterization of the Empirical Mean and Median

The empirical mean is the center of x_1, ..., x_n, where center is defined to minimize squared distance.

The empirical median is the center of x_1, ..., x_n, where center is defined to minimize absolute distance.

Empirical Distribution Calculations in R

Suppose the vector $(x_1, \ldots, x_n)$ has been made an R vector, for example, by

```
x <- c(0.03, 0.04, 0.05, 0.49, 0.50,
    0.59, 0.66, 0.72, 0.83, 1.17)
```

Then

```
mean(x)
```

calculates the empirical mean for these data and

```
median(x)
```

calculates the empirical median.

Empirical Distribution Calculations in R (cont.)

Furthermore, the `mean` function can be used to calculate other empirical expectations, for example,

```
xbar <- mean(x)
mean((x - xbar)^2)
```

calculates the empirical variance, as does the one-liner

```
mean((x - mean(x))^2)
```

Empirical Distribution Calculations in R (cont.)

```
bigf <- ecdf(x)
```

calculates the empirical distribution function (the "c" in `ecdf` is for "cumulative" because non-theoretical people call DF "cumulative distribution functions"). The result is a function that can be evaluated at any real number.

```
bigf(0)
bigf(0.5)
bigf(1)
bigf(1.5)
```

and so forth.

Empirical Distribution Calculations in R (cont.)

The empirical DF can also be plotted by

```
plot(bigf)
```

or by the one-liner

```
plot(ecdf(x))
```

Empirical Distribution Calculations in R (cont.)

R also has a function `quantile` that calculates quantiles of the empirical distribution. As we mentioned there is no widely accepted notion of the best way to calculate quantiles. The definition we gave is simple and theoretically correct, but arguments can be given for other notions and the `quantile` function can calculate no less than nine different notions of "quantile".

```
quantile(x, type = 1)
```

calculates a bunch of quantiles. Other quantiles can be specified

```
quantile(x, probs = 1 / 3, type = 1)
```

Little x to Big X

We now do something really tricky.

So far we have just been reviewing finite probability spaces. The numbers x_1, ..., x_n are just numbers.

Now we want to make the numbers X_1, ..., X_n that determine the empirical distribution IID random variables.

In one sense the change is trivial: capitalize all the x's you see.

In another sense the change is profound: now all the thingummies of interest — mean, variance, other moments, median, quantiles, and DF of the empirical distribution — are random variables.

Little x to Big X (cont.)

For example

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

(the mean of the empirical distribution) is a random variable. What is the distribution of this random variable? It is determined somehow by the distribution of the X_i.

When the distribution of $\overline{X}_n$ is not a brand-name distribution but the distribution of

$$n\overline{X}_n = \sum_{i=1}^{n} X_i$$

is a brand name distribution, then we refer to that.

Sampling Distribution of the Empirical Mean

The distribution of $n\overline{X}_n$ is given by what the brand-name distribution handout calls "addition rules".

If each X_i is Ber(p), then $n\overline{X}_n$ is Bin(n, p).

If each X_i is Geo(p), then $n\overline{X}_n$ is NegBin(n, p).

If each X_i is Poi(μ), then $n\overline{X}_n$ is Poi($n\mu$).

If each X_i is Exp(λ), then $n\overline{X}_n$ is Gam(n, λ).

If each X_i is $\mathcal{N}(\mu, \sigma^2)$, then $n\overline{X}_n$ is $\mathcal{N}(n\mu, n\sigma^2)$.

Sampling Distribution of the Empirical Mean (cont.)

In the latter two cases, we can apply the change-of-variable theorem to the linear transformation $y = x/n$ obtaining

$$f_Y(y) = n f_X(ny)$$

If each X_i is Exp(λ), then $\overline{X}_n$ is Gam($n, n\lambda$).

If each X_i is $\mathcal{N}(\mu, \sigma^2)$, then $\overline{X}_n$ is $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Sampling Distribution of the Empirical Mean (cont.)

For most distributions of of the X_i we cannot calculate the exact sampling distribution of $n\overline{X}_n$ or of $\overline{X}_n$. The central limit theorem (CLT), however, gives an approximation of the sampling distribution when n is large.

If each X_i has mean μ and variance σ^2, then $\overline{X}_n$ is approximately $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

The CLT is not applicable if the X_i do not have finite variance.

Sampling Distributions

The same game can be played with any of the other quantities, the empirical median, for example.

Much more can be said about the empirical mean, because we have the addition rules to work with. The distribution of the empirical median is not brand-name unless the X_i are Unif$(0, 1)$ and n is odd. There is a large n approximation, but the argument is long and complicated. We will do both, but not right away.

Sampling Distributions (cont.)

The important point to understand for now is that any random variable has a distribution (whether we can name it or otherwise describe it), hence these quantities related to the empirical distribution have probability distributions — called their *sampling distributions* — and we can sometimes describe them exactly, sometimes give large n approximations, and sometimes not even that. But they always exist, whether we can describe them or not, and we can refer to them in theoretical arguments.

Sampling Distributions (cont.)

Why the "sample" in "sampling distribution"?

Suppose X_1, ..., X_n are a sample with replacement from a finite population. Then we say the distribution of each X_i is the *population distribution*, and we say X_1, ..., X_n are a *random sample* from this population, and we say the distribution of $\overline{X}_n$ is its *sampling distribution* because its randomness comes from X_1, ..., X_n being a random sample.

This is the story that introduces sampling distributions in most intro stats courses. It is also the language that statisticians use in talking to people who haven't had a theory course like this one.

Sampling Distributions (cont.)

This language becomes only a vague metaphor when X_1, ..., X_n are IID but their distribution does not have a finite sample space, so they cannot be considered — strictly speaking — a sample from a finite population.

They can be considered a sample from an infinite population in a vague metaphorical way, but when we try to formalize this notion we cannot. Strictly speaking it is nonsense.

And strictly speaking, the "sampling" in "sampling distribution" is redundant. The "sampling distribution" of $\overline{X}_n$ is the distribution of $\overline{X}_n$. It is a random variable, hence it has a probability distribution. Its probability distribution doesn't need the adjective "sampling" attached to it any more than any other probability distribution does (i. e., not at all).

Sampling Distributions (cont.)

So why do statisticians, who are serious people, persist in using this rather silly language? The phrase "sampling distribution" alerts the listener that we are not talking about the "population distribution" and the distribution of $\overline{X}_n$ or $\widetilde{X}_n$ (or whatever quantity related to the empirical distribution is under discussion) is not the same as the distribution of each X_i.

Of course, no one theoretically sophisticated (like all of you) would think for a second that the distribution of $\overline{X}_n$ is the same as the distribution of the X_i, but — probability being hard for less sophisticated audiences — the stress in "sampling distribution" — redundant though it may be — is perhaps useful.

Chi-Square Distribution

Recall that for any real number $\nu > 0$ the *chi-square distribution having* ν *degrees of freedom*, abbreviated $\text{chi}^2(\nu)$, is another name for the $\Gamma\left(\frac{\nu}{2}, \frac{1}{2}\right)$ distribution.

Student's T Distribution

Now we come to a new brand name distribution whose name is the single letter t (not very good terminology). It is sometimes called "Student's t distribution" because it was invented by W. S. Gosset who published under the pseudonym "Student".

Suppose Z and Y are independent random variables

$$\begin{aligned} Z &\sim \mathcal{N}(0,1) \\ Y &\sim \text{chi}^2(\nu) \end{aligned}$$

then

$$T = \frac{Z}{\sqrt{Y/\nu}}$$

is said to have *Student's t distribution with ν degrees of freedom*, abbreviated $t(\nu)$.

Student's T Distribution (cont.)

The PDF of the $t(\nu)$ distribution is

$$f_\nu(x) = \frac{1}{\sqrt{\nu\pi}} \cdot \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \cdot \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}, \qquad -\infty < x < +\infty$$

because $\Gamma(\frac{1}{2}) = \sqrt{\pi}$,

$$\frac{1}{\sqrt{\nu\pi}} \cdot \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} = \frac{1}{\sqrt{\nu}} \cdot \frac{1}{B(\frac{\nu}{2}, \frac{1}{2})}$$

where the beta function $B(\frac{\nu}{2}, \frac{1}{2})$ is the normalizing constant of the beta distribution defined in the brand name distributions handout.

Student's T Distribution (cont.)

The joint distribution of Z and Y in the definition is

$$f(z, y) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \frac{\left(\frac{1}{2}\right)^{\nu/2}}{\Gamma(\nu/2)} y^{\nu/2-1} e^{-y/2}$$

Make the change of variables $t = z/\sqrt{y/\nu}$ and $u = y$, which has inverse transformation

$$\begin{aligned} z &= t\sqrt{u/\nu} \\ y &= u \end{aligned}$$

and Jacobian

$$\begin{vmatrix} \sqrt{u/\nu} & t/2\sqrt{u\nu} \\ 0 & 1 \end{vmatrix} = \sqrt{u/\nu}$$

Student's T Distribution (cont.)

The joint distribution of T and U given by the multivariate change of variable formula is

$$
\begin{aligned}
f(t, u) &= \frac{1}{\sqrt{2\pi}} e^{-(t\sqrt{u/\nu})^2/2} \frac{\left(\frac{1}{2}\right)^{\nu/2}}{\Gamma(\nu/2)} u^{\nu/2-1} e^{-u/2} \cdot \sqrt{u/\nu} \\
&= \frac{1}{\sqrt{2\pi}} \frac{\left(\frac{1}{2}\right)^{\nu/2}}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu}} u^{\nu/2-1/2} \exp\left\{-\left(1 + \frac{t^2}{\nu}\right)\frac{u}{2}\right\}
\end{aligned}
$$

Thought of as a function of u for fixed t, this is proportional to a gamma density with shape parameter $(\nu + 1)/2$ and rate parameter $\frac{1}{2}(1 + \frac{t^2}{\nu})$.

Student's T Distribution (cont.)

The "recognize the unnormalized density trick" which is equivalent to using the "theorem" for the gamma distribution allows us to integrate out u getting the marginal of t

$$f(t) = \frac{1}{\sqrt{2\pi}} \cdot \frac{\left(\frac{1}{2}\right)^{\nu/2}}{\Gamma(\nu/2)} \cdot \frac{1}{\sqrt{\nu}} \cdot \frac{\Gamma(\frac{\nu+1}{2})}{[\frac{1}{2}(1 + \frac{t^2}{\nu})]^{(\nu+1)/2}}$$

which, after changing t to x, simplifies to the form given on slide 42.

Student's T Distribution: Moments

The t distribution is symmetric about zero, hence the mean is zero if the mean exists. Hence central moments are equal to ordinary moments. Hence every odd ordinary moment is zero if it exists.

For the $t(\nu)$ distribution and $k > 0$, the ordinary moment $E(|X|^k)$ exists if and only if $k < \nu$.

Student's T Distribution: Moments (cont.)

The PDF is bounded, so the question of whether moments exist only involves behavior of the PDF at $\pm\infty$. Since the t distribution is symmetric about zero, we only need to check the behavior at $+\infty$. When does

$$\int_0^\infty x^k f_\nu(x)\,dx$$

exist? Since

$$\lim_{x\to\infty} \frac{x^k f_\nu(x)}{x^\alpha} \to c$$

when $\alpha = k - (\nu + 1)$. The comparison theorem (5101 Slide 9, Deck 4) says the integral exists if and only if

$$k - (\nu + 1) = \alpha < -1$$

which is equivalent to $k < \nu$.

Student's T Distribution: Moments (cont.)

If X has the $t(\nu)$ distribution and $\nu > 1$, then

$$E(X) = 0$$

Otherwise the mean does not exist. (Proof: symmetry.)

If X has the $t(\nu)$ distribution and $\nu > 2$, then

$$\text{var}(X) = \frac{\nu}{\nu - 2}$$

Otherwise the variance does not exist. (Proof: homework.)

Student's T Distribution and Cauchy Distribution

Plugging in $\nu = 1$ into the formula for the PDF of the $t(\nu)$ distribution on slide 42 gives the PDF of the standard Cauchy distribution. In short $t(1) = \text{Cauchy}(0, 1)$.

Hence if Z_1 and Z_2 are independent $\mathcal{N}(0, 1)$ random variables, then

$$T = \frac{Z_1}{Z_2}$$

has the Cauchy$(0, 1)$ distribution.

Student's T Distribution and Normal Distribution

If Y_ν is chi$^2(\nu) = \text{Gam}(\frac{\nu}{2}, \frac{1}{2})$, then $U_\nu = Y_\nu / \nu$ is $\text{Gam}(\frac{\nu}{2}, \frac{\nu}{2})$, and

$$E(U_\nu) = 1$$
$$\text{var}(U_\nu) = \frac{2}{\nu}$$

Hence

$$U_\nu \xrightarrow{P} 1, \qquad \text{as } \nu \to \infty$$

by Chebyshev's inequality. Hence if Z is a standard normal random variable independent of Y_ν

$$\frac{Z}{\sqrt{Y_\nu/\nu}} \xrightarrow{\mathcal{D}} Z, \qquad \text{as } \nu \to \infty$$

by Slutsky's theorem. In short, the $t(\nu)$ distribution converges to the $\mathcal{N}(0,1)$ distribution as $\nu \to \infty$.

Snedecor's F Distribution

If X and Y are independent random variables and

$$X \sim \text{chi}^2(\nu_1)$$
$$Y \sim \text{chi}^2(\nu_2)$$

then

$$W = \frac{X/\nu_1}{Y/\nu_2}$$

has the F *distribution with* ν_1 *numerator degrees of freedom and* ν_2 *denominator degrees of freedom*.

Snedecor's F Distribution (cont.)

The "F" is for R. A. Fisher, who introduced a function of this random variable into statistical inference. This particular random variable was introduced by G. Snedecor. Hardly anyone knows this history or uses the eponyms.

This is our second brand-name distribution whose name is a single letter. It is abbreviated $F(\nu_1, \nu_2)$.

Snedecor's F Distribution (cont.)

The theorem on slides 130–139, 5101 Deck 3 says that if X and Y are independent random variables and

$$
\begin{aligned}
X &\sim \text{Gam}(\alpha_1, \lambda) \\
Y &\sim \text{Gam}(\alpha_2, \lambda)
\end{aligned}
$$

then

$$
V = \frac{X}{X + Y}
$$

has the Beta(α_1, α_2) distribution.

Snedecor's F Distribution (cont.)

Hence, if X and Y are independent random variables and

$$X \sim \text{chi}^2(\nu_1)$$
$$Y \sim \text{chi}^2(\nu_2)$$

then

$$V = \frac{X}{X + Y}$$

has the Beta$(\frac{\nu_1}{2}, \frac{\nu_2}{2})$ distribution.

Snedecor's F Distribution (cont.)

Since

$$\frac{X}{Y} = \frac{V}{1-V}$$

we have

$$W = \frac{\nu_2}{\nu_1} \cdot \frac{V}{1-V}$$

and

$$V = \frac{\nu_1 W/\nu_2}{1 + \nu_1 W/\nu_2}$$

This gives the relationship between the $F(\nu_1, \nu_2)$ distribution of W and the Beta$(\frac{\nu_1}{2}, \frac{\nu_2}{2})$ distribution of V.

Snedecor's F Distribution (cont.)

The PDF of the F distribution can be derived from the PDF of the beta distribution using the change-of-variable formula. It is given in the brand name distributions handout, but is not very useful.

If one wants moments of the F distribution, for example,

$$E(W) = \frac{\nu}{\nu - 2}$$

when $\nu > 2$, write W as a function of V and calculate the moment that way.

Snedecor's F Distribution (cont.)

The same argument used to show

$$t(\nu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \qquad \text{as } \nu \to \infty$$

shows

$$F(\nu_1, \nu_2) \xrightarrow{P} 1, \qquad \text{as } \nu_1 \to \infty \text{ and } \nu_2 \to \infty$$

So an F random variable is close to 1 when both degrees of freedom are large.

Sampling Distributions for Normal Populations

Suppose X_1, ..., X_n are IID $\mathcal{N}(\mu, \sigma^2)$ and

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$V_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

are the mean and variance of the empirical distribution. Then $\overline{X}_n$ and V_n are independent random variables and

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{nV_n}{\sigma^2} \sim \text{chi}^2(n-1)$$

Sampling Distributions for Normal Populations (cont.)

It is traditional to name the distribution of

$$\frac{nV_n}{\sigma^2} = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

rather than of V_n itself. But, of course, if

$$\frac{nV_n}{\sigma^2} \sim \text{chi}^2(n-1)$$

then

$$V_n \sim \text{Gam}\left(\frac{n-1}{2}, \frac{n}{2\sigma^2}\right)$$

by the change-of-variable theorem.

Sampling Distributions for Normal Populations (cont.)

Strictly speaking, the "populations" in the heading should be in scare quotes, because infinite populations are vague metaphorical nonsense.

Less pedantically, it is important to remember that the theorem on slide 58 has no analog for non-normal populations.

In general, $\overline{X}_n$ and V_n are not independent.

In general, the sampling distribution of $\overline{X}_n$ is not exactly $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, although it is approximately so when n is large.

In general, the sampling distribution of V_n is not exactly gamma.

Empirical Variance and Sample Variance

Those who have been exposed to an introductory statistic course may be wondering why we keep saying "empirical mean" rather than "sample mean" which everyone else says. The answer is that the "empirical variance" V_n is not what everyone else calls the "sample variance".

In general, we do not know the distribution of V_n. It is not brand name and is hard or impossible to describe explicitly.

However we always have

$$E(V_n) = \frac{n-1}{n} \cdot \sigma^2$$

Empirical Variance and Sample Variance (cont.)

Define

$$V_n^* = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

where $\mu = E(X_i)$. Then $E(V_n^*) = \sigma^2$, because $E\{(X_i - \mu)^2\} = \sigma^2$.

The empirical analog of the mean square error formula (derived on slide 7) is

$$E_n\{(X - a)^2\} = \text{var}_n(X) + (a - \overline{X}_n)^2$$

and plugging in μ for a gives

$$V_n^* = E_n\{(X - \mu)^2\} = \text{var}_n(X) + (\mu - \overline{X}_n)^2 = V_n + (\mu - \overline{X}_n)^2$$

Empirical Variance and Sample Variance (cont.)

But since $E(\overline{X}_n) = \mu$ (5101, Slide 89, Deck 2)

$$E\{(\mu - \overline{X}_n)^2\} = \text{var}(\overline{X}_n)$$

In summary,

$$E(V_n^*) = E(V_n) + \text{var}(\overline{X}_n)$$

and we know $\text{var}(\overline{X}_n) = \sigma^2 / n$ (5101, Slide 89, Deck 2), so

$$E(V_n) = E(V_n^*) - \text{var}(\overline{X}_n) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \cdot \sigma^2$$

Empirical Variance and Sample Variance (cont.)

The factor $(n-1)/n$ is deemed to be unsightly, so

$$S_n^2 = \frac{n}{n-1} \cdot V_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

which has the simpler property

$$E(S_n^2) = \sigma^2$$

is usually called the *sample variance*, and S_n is usually called the *sample standard deviation*.

In cookbook applied statistics the fact that these are not the variance and standard deviation of the empirical distribution does no harm. But it does mess up the theory. So we do not take S_n^2 as being the obvious quantity to study and look at V_n too.

Sampling Distributions for Normal Populations (cont.)

We now prove the theorem stated on slide 58.

The random vector $(\overline{X}_n, X_1 - \overline{X}_n, \ldots, X_n - \overline{X}_n)$, being a linear function of a multivariate normal, is multivariate normal.

We claim the first component $\overline{X}_n$ is independent of the other components $X_i - \overline{X}_n$, $i = 1$, ..., n. Since uncorrelated implies independent for multivariate normal (5101, Deck 5, Slides 129–134), it is enough to verify

$$\text{cov}(\overline{X}_n, X_i - \overline{X}_n) = 0$$

Sampling Distributions for Normal Populations (cont.)

$$
\begin{aligned}
\operatorname{cov}(\overline{X}_n, X_i - \overline{X}_n) &= \operatorname{cov}(\overline{X}_n, X_i) - \operatorname{var}(\overline{X}_n) \\
&= \operatorname{cov}(\overline{X}_n, X_i) - \frac{\sigma^2}{n} \\
&= \operatorname{cov}\left(\frac{1}{n}\sum_{j=1}^{n} X_j, X_i\right) - \frac{\sigma^2}{n} \\
&= \frac{1}{n}\sum_{j=1}^{n} \operatorname{cov}(X_j, X_i) - \frac{\sigma^2}{n} \\
&= 0
\end{aligned}
$$

because $\operatorname{cov}(X_j, X_i) = 0$ when $i \neq j$ and because $\operatorname{cov}(X_i, X_i) = \operatorname{var}(X_i) = \sigma^2$.

Sampling Distributions for Normal Populations (cont.)

That finishes the proof that $\overline{X}_n$ and V_n are independent random variables, because V_n is a function of $X_i - \overline{X}_n$, $i = 1, \ldots, n$.

That

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

we already knew. It comes from the addition rule for the normal distribution.

Establishing the sampling distribution of V_n is more complicated.

Orthonormal Bases and Orthogonal Matrices

A set of vectors $\mathcal{U}$ is *orthonormal* if each has length one

$$\mathbf{u}^T\mathbf{u} = 1, \qquad \mathbf{u} \in \mathcal{U}$$

and each pair is orthogonal

$$\mathbf{u}^T\mathbf{v} = 0, \qquad \mathbf{u}, \mathbf{v} \in \mathcal{U} \text{ and } \mathbf{u} \neq \mathbf{v}$$

An orthonormal set of d vectors in d-dimensional space is called an *orthonormal basis* (plural *orthonormal bases*, pronounced like "base ease").

Orthonormal Bases and Orthogonal Matrices (cont.)

A square matrix whose columns form an orthonormal basis is called *orthogonal*.

If $\mathbf{O}$ is orthogonal, then the orthonormality property expressed in matrix notation is

$$\mathbf{O}^T \mathbf{O} = \mathbf{I}$$

where $\mathbf{I}$ is the identity matrix. This implies $\mathbf{O}^T = \mathbf{O}^{-1}$ and

$$\mathbf{O}\mathbf{O}^T = \mathbf{I}$$

Hence the rows of $\mathbf{O}$ also form an orthonormal basis.

Orthonormal matrices have appeared before in the spectral decomposition (5101 Deck 5, Slides 103–110).

Orthonormal Bases and Orthogonal Matrices (cont.)

It is a theorem of linear algebra, which we shall not prove, that any orthonormal set of vectors can be extended to an orthonormal basis (the Gram-Schmidt orthogonalization process can be used to do this).

The unit vector

$$\mathbf{u} = \frac{1}{\sqrt{n}}(1, 1, \ldots, 1)$$

all of whose components are the same forms an orthonormal set $\{\mathbf{u}\}$ of size one. Hence there exists an orthogonal matrix $\mathbf{O}$ whose first column is $\mathbf{u}$.

Sampling Distributions for Normal Populations (cont.)

Any orthogonal matrix $\mathbf{O}$ maps standard normal random vectors to standard normal random vectors. If $\mathbf{Z}$ is standard normal and $\mathbf{Y} = \mathbf{O}^T \mathbf{Z}$, then

$$
\begin{aligned}
E(\mathbf{Y}) &= \mathbf{O}^T E(\mathbf{Z}) = 0 \\
\operatorname{var}(\mathbf{Y}) &= \mathbf{O}^T \operatorname{var}(\mathbf{Z}) \mathbf{O} = \mathbf{O}^T \mathbf{O} = \mathbf{I}
\end{aligned}
$$

Sampling Distributions for Normal Populations (cont.)

Also

$$
\begin{aligned}
\sum_{i=1}^{n} Y_i^2 &= \|\mathbf{Y}\|^2 \\
&= \mathbf{Y}^T \mathbf{Y} \\
&= \mathbf{Z}^T \mathbf{O}^T \mathbf{O} \mathbf{Z} \\
&= \mathbf{Z}^T \mathbf{Z} \\
&= \sum_{i=1}^{n} Z_i^2
\end{aligned}
$$

Sampling Distributions for Normal Populations (cont.)

In the particular case where $\mathbf{u}$ is the first column of $\mathbf{O}$

$$\sum_{i=1}^{n} Y_i^2 = Y_1^2 + \sum_{i=2}^{n} Y_i^2$$
$$= n\overline{Z}_n^2 + \sum_{i=2}^{n} Y_i^2$$

because

$$\mathbf{u}^T\mathbf{Z} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i = \sqrt{n}\,\overline{Z}_n$$

Sampling Distributions for Normal Populations (cont.)

Hence

$$
\begin{aligned}
\sum_{i=2}^{n} Y_i^2 &= \sum_{i=1}^{n} Y_i^2 - n \overline{Z}_n^2 \\
&= \sum_{i=1}^{n} Z_i^2 - n \overline{Z}_n^2 \\
&= n \left(\frac{1}{n} \sum_{i=1}^{n} Z_i^2 - \overline{Z}_n^2 \right) \\
&= n \operatorname{var}_n(Z)
\end{aligned}
$$

This establishes the theorem in the special case $\mu = 0$ and $\sigma^2 = 1$ because the components of $\mathbf{Y}$ are IID standard normal, hence n times the empirical variance of Z_1, ..., Z_n has the chi-square distribution with $n - 1$ degrees of freedom.

Sampling Distributions for Normal Populations (cont.)

To finish the proof of the theorem, notice that if X_1, ..., X_n are IID $\mathcal{N}(\mu, \sigma^2)$, then

$$Z_i = \frac{X_i - \mu}{\sigma}, \qquad i = 1, \ldots, n$$

are IID standard normal. Hence

$$n \operatorname{var}_n(Z) = \frac{n \operatorname{var}_n(X)}{\sigma^2} = \frac{n V_n}{\sigma^2}$$

has the chi-square distribution with $n - 1$ degrees of freedom. That finishes the proof of the theorem stated on slide 58.

Sampling Distributions for Normal Populations (cont.)

The theorem can be stated with S_n^2 replacing V_n. If X_1, ..., X_n are IID $\mathcal{N}(\mu, \sigma^2)$ and

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

then $\overline{X}_n$ and S_n^2 are independent random variables and

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \text{chi}^2(n-1)$$

Sampling Distributions for Normal Populations (cont.)

An important consequence uses the theorem as restated using S_n^2 and the definition of a $t(n-1)$ random variable.

If X_1, ..., X_n are IID $\mathcal{N}(\mu, \sigma^2)$, then

$$\frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1)$$
$$\frac{(n-1)S_n^2}{\sigma^2} \sim \text{chi}^2(n-1)$$

Hence

$$T = \frac{(\overline{X}_n - \mu)/\sigma/\sqrt{n}}{\sqrt{[(n-1)S_n^2/\sigma^2]/(n-1)}} = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$$

has the $t(n-1)$ distribution.

Asymptotic Sampling Distributions

When the data X_1, ..., X_n are IID from a distribution that is not normal, we have no result like the theorem just discussed for the normal distribution. Even when the data are IID normal, we have no exact sampling distribution for moments other than the mean and variance. We have to make do with asymptotic, large n, approximate results.

Asymptotic Sampling Distributions (cont.)

The ordinary and central moments of the distribution of the data were defined on 5101 deck 3, slides 153–154. The ordinary moments, if they exist, are denoted

$$\alpha_k = E(X_i^k)$$

(they are the same for all i because the data are IID). The first ordinary moment is the mean $\mu = \alpha_1$. The central moments, if they exist, are denoted

$$\mu_k = E\{(X_i - \mu)^k\}$$

(they are the same for all i because the data are IID). The first central moment is always zero $\mu_1 = 0$. The second central moment is the variance $\mu_2 = \sigma^2$.

Asymptotic Sampling Distributions (cont.)

The ordinary and central moments of the empirical distribution are defined in the same way. The ordinary moments are denoted

$$A_{k,n} = E_n(X^k) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

The first ordinary moment is the empirical mean $\overline{X}_n = A_{1,n}$. The central moments are denoted

$$M_{k,n} = E_n\{(X - \overline{X}_n)^k\} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^k$$

The first central moment is always zero $M_{1,n} = 0$. The second central moment is the empirical variance $M_{2,n} = V_n$.

Asymptotic Sampling Distributions (cont.)

The asymptotic joint distribution of the ordinary empirical moments was done on 5101 deck 6, slides 89–90 although we hadn't introduced the empirical distribution yet so didn't describe it this way.

Asymptotic Sampling Distributions (cont.)

Define random vectors

$$\mathbf{Y}_i = \begin{pmatrix} X_i \\ X_i^2 \\ \vdots \\ X_i^k \end{pmatrix}$$

Then

$$\overline{\mathbf{Y}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i = \begin{pmatrix} A_{1,n} \\ A_{2,n} \\ \vdots \\ A_{k,n} \end{pmatrix}$$

Asymptotic Sampling Distributions (cont.)

$$
E(\mathbf{Y}_i) = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix}
$$

$$
\operatorname{var}(\mathbf{Y}_i) = \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1 \alpha_2 & \cdots & \alpha_{k+1} - \alpha_1 \alpha_k \\ \alpha_3 - \alpha_1 \alpha_2 & \alpha_4 - \alpha_2^2 & \cdots & \alpha_{k+2} - \alpha_2 \alpha_k \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{k+1} - \alpha_1 \alpha_k & \alpha_{k+2} - \alpha_2 \alpha_k & \cdots & \alpha_{2k} - \alpha_k^2 \end{pmatrix}
$$

(they are the same for all i because the data are IID). Details of the variance calculation are on 5101 deck 6, slide 89.

Asymptotic Sampling Distributions (cont.)

Write

$$\begin{aligned} E(\mathbf{Y}_i) &= \boldsymbol{\mu}_{\text{ordinary}} \\ \text{var}(\mathbf{Y}_i) &= \mathbf{M}_{\text{ordinary}} \end{aligned}$$

($\boldsymbol{\mu}_{\text{ordinary}}$ is a vector and $\mathbf{M}_{\text{ordinary}}$ is a matrix). Then the multivariate CLT (5101 deck 6, slides 74–86) says

$$\overline{\mathbf{Y}}_n \approx \mathcal{N}\left(\boldsymbol{\mu}_{\text{ordinary}}, \frac{\mathbf{M}_{\text{ordinary}}}{n}\right)$$

Since the components of $\mathbf{Y}_n$ are the empirical ordinary moments up to order k, this gives the asymptotic (large n, approximate) joint distribution of the empirical ordinary moments up to order k. Since $\mathbf{M}_{\text{ordinary}}$ contains population moments up to order $2k$, we need to assume those exist.

Asymptotic Sampling Distributions (cont.)

All of this about empirical ordinary moments is simple — a straightforward application of the multivariate CLT — compared to the analogous theory for empirical central moments. The problem is that

$$M_{k,n} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^k$$

is not an empirical mean of the form

$$E_n\{g(X)\} = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

for any function g.

Asymptotic Sampling Distributions (cont.)

We would have a simple theory, analogous to the theory for empirical ordinary moments if we studied instead

$$M^*_{k,n} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^k$$

which are empirical moments but are not functions of data only so not as interesting.

It turns out that the asymptotic joint distribution of the $M^*_{k,n}$ is theoretically useful as a step on the way to the asymptotic joint distribution of the $M_{k,n}$, so let's do it.

Asymptotic Sampling Distributions (cont.)

Define random vectors

$$\mathbf{Z}_i^* = \begin{pmatrix} X_i - \mu \\ (X_i - \mu)^2 \\ \vdots \\ (X_i - \mu)^k \end{pmatrix}$$

Then

$$\overline{\mathbf{Z}}_n^* = \begin{pmatrix} M_{1,n}^* \\ M_{2,n}^* \\ \vdots \\ M_{k,n}^* \end{pmatrix}$$

Asymptotic Sampling Distributions (cont.)

$$
E(\mathbf{Z}_i^*) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}
$$

$$
\operatorname{var}(\mathbf{Z}_i^*) = \begin{pmatrix} \mu_2 - \mu_1^2 & \mu_3 - \mu_1 \mu_2 & \cdots & \mu_{k+1} - \mu_1 \mu_k \\ \mu_3 - \mu_1 \mu_2 & \mu_4 - \mu_2^2 & \cdots & \mu_{k+2} - \mu_2 \mu_k \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{k+1} - \mu_1 \mu_k & \mu_{k+2} - \mu_2 \mu_k & \cdots & \mu_{2k} - \mu_k^2 \end{pmatrix}
$$

(they are the same for all i because the data are IID). The variance calculation follows from the one for ordinary moments because central moments of X_i are ordinary moments of $X_i - \mu$.

Asymptotic Sampling Distributions (cont.)

Write

$$E(\mathbf{Z}_i^*) = \boldsymbol{\mu}_{\text{central}}$$
$$\text{var}(\mathbf{Z}_i^*) = \mathbf{M}_{\text{central}}$$

($\boldsymbol{\mu}_{\text{central}}$ is a vector and $\mathbf{M}_{\text{central}}$ is a matrix). Then the multivariate CLT (5101 deck 6, slides 74–86) says

$$\overline{\mathbf{Z}}_n^* \approx \mathcal{N}\left(\boldsymbol{\mu}_{\text{central}}, \frac{\mathbf{M}_{\text{central}}}{n}\right)$$

Since the components of $\mathbf{Z}_n^*$ are the $M_{i,n}^*$ up to order k, this gives the asymptotic (large n, approximate) joint distribution of the $M_{i,n}^*$ up to order k. Since $\mathbf{M}_{\text{central}}$ contains population moments up to order $2k$, we need to assume those exist.

Asymptotic Sampling Distributions (cont.)

These theorems imply the laws of large numbers (LLN)

$$A_{k,n} \xrightarrow{P} \alpha_k$$
$$M^*_{k,n} \xrightarrow{P} \mu_k$$

for each k, but these LLN actually hold under the weaker conditions that the population moments on the right-hand side exist.

The CLT for $A_{k,n}$ requires population moments up to order $2k$. The LLN for $A_{k,n}$ requires population moments up to order k. Similarly for $M^*_{k,n}$.

Asymptotic Sampling Distributions (cont.)

By the binomial theorem

$$
\begin{aligned}
M_{k,n} &= \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^k \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{k} \binom{k}{j} (-1)^j (\overline{X}_n - \mu)^j (X_i - \mu)^{k-j} \\
&= \sum_{j=0}^{k} \binom{k}{j} (-1)^j (\overline{X}_n - \mu)^j \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^{k-j} \\
&= \sum_{j=0}^{k} \binom{k}{j} (-1)^j (\overline{X}_n - \mu)^j M^*_{k-j,n}
\end{aligned}
$$

Asymptotic Sampling Distributions (cont.)

By the LLN

$$\overline{X}_n \xrightarrow{P} \mu$$

so by the continuous mapping theorem

$$(\overline{X}_n - \mu)^j \xrightarrow{P} 0$$

for any positive integer j. Hence by Slutsky's theorem

$$\binom{k}{j}(-1)^j (\overline{X}_n - \mu)^j M^*_{k-j,n} \xrightarrow{P} 0$$

for any positive integer j. Hence by another application of Slutsky's theorem

$$M_{k,n} \xrightarrow{P} \mu_k$$

Asymptotic Sampling Distributions (cont.)

Define random vectors

$$\mathbf{Z}_i = \begin{pmatrix} X_i - \overline{X}_n \\ (X_i - \overline{X}_n)^2 \\ \vdots \\ (X_i - \overline{X}_n)^k \end{pmatrix}$$

Then

$$\overline{\mathbf{Z}}_n = \begin{pmatrix} M_{1,n} \\ M_{2,n} \\ \vdots \\ M_{k,n} \end{pmatrix}$$

Asymptotic Sampling Distributions (cont.)

Since convergence in probability to a constant of random vectors is merely convergence in probability to a constant of each component (5101, deck 6, slides 68–73), we can write these univariate LLN as multivariate LLN

$$\overline{\mathbf{Z}}_n^* \xrightarrow{P} \boldsymbol{\mu}_{\text{central}}$$

$$\overline{\mathbf{Z}}_n \xrightarrow{P} \boldsymbol{\mu}_{\text{central}}$$

Asymptotic Sampling Distributions (cont.)

Up to now we used the "sloppy" version of the multivariate CLT and it did no harm because we went immediately to the conclusion. Now we want to apply Slutsky's theorem, so we need the careful pedantically correct version. The sloppy version was

$$\overline{\mathbf{Z}}_n^* \approx \mathcal{N}\left(\boldsymbol{\mu}_{\text{central}}, \frac{\mathbf{M}_{\text{central}}}{n}\right)$$

The careful version is

$$\sqrt{n}\big(\overline{\mathbf{Z}}_n^* - \boldsymbol{\mu}_{\text{central}}\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M}_{\text{central}})$$

The careful version has no n in the limit (right-hand side), as must be the case for any limit as $n \to \infty$. The sloppy version does have an n on the right-hand side, which consequently cannot be a mathematical limit.

Asymptotic Sampling Distributions (cont.)

$$\begin{aligned}
&\sqrt{n}(M_{k,n} - \mu_k) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[(X_i - \overline{X}_n)^k - \mu_k \right] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\sum_{j=0}^{k} \binom{k}{j} (-1)^j (\overline{X}_n - \mu)^j (X_i - \mu)^{k-j} - \mu_k \right] \\
&= \sqrt{n}(M_{k,n}^* - \mu_k) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{k} \binom{k}{j} (-1)^j (\overline{X}_n - \mu)^j (X_i - \mu)^{k-j} \\
&= \sqrt{n}(M_{k,n}^* - \mu_k) + \sum_{j=1}^{k} \binom{k}{j} (-1)^j \sqrt{n} (\overline{X}_n - \mu)^j M_{k-j,n}^*
\end{aligned}$$

Asymptotic Sampling Distributions (cont.)

By the CLT

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} U$$

where $U \sim \mathcal{N}(0, \sigma^2)$. Hence by the continuous mapping theorem

$$n^{j/2}(\overline{X}_n - \mu)^j \xrightarrow{\mathcal{D}} U^j$$

but by Slutsky's theorem

$$\sqrt{n}(\overline{X}_n - \mu)^j \xrightarrow{\mathcal{D}} 0, \qquad j = 2, 3 \ldots$$

Asymptotic Sampling Distributions (cont.)

Hence only the $j = 0$ and $j = 1$ terms on slide 96 do not converge in probability to zero, that is,

$$\sqrt{n}(M_{k,n} - \mu_k) = \sqrt{n}(M^*_{k,n} - \mu_k) - k\sqrt{n}(\overline{X}_n - \mu)M^*_{k-1,n} + o_p(1)$$

where $o_p(1)$ means terms that converge in probability to zero.

By Slutsky's theorem this converges to

$$W - k\mu_{k-1}U$$

where the bivariate random vector (U, W) is multivariate normal with mean vector zero and variance matrix

$$\mathbf{M} = \text{var}\begin{pmatrix} X_i - \mu \\ (X_i - \mu)^k \end{pmatrix} = \begin{pmatrix} \mu_2 & \mu_{k+1} \\ \mu_{k+1} & \mu_{2k} - \mu_k^2 \end{pmatrix}$$

Asymptotic Sampling Distributions (cont.)

Apply the multivariate delta method, which in this case says that the distribution of

$$W - k\mu_{k-1} U$$

is univariate normal with mean zero and variance

$$\begin{aligned} &\begin{pmatrix} -k\mu_{k-1} & 1 \end{pmatrix} \begin{pmatrix} \mu_2 & \mu_{k+1} \\ \mu_{k+1} & \mu_{2k} - \mu_k^2 \end{pmatrix} \begin{pmatrix} -k\mu_{k-1} \\ 1 \end{pmatrix} \\ &\qquad = \mu_{2k} - \mu_k^2 - 2k\mu_{k-1}\mu_{k+1} + k^2\mu_{k-1}^2\mu_2 \end{aligned}$$

Asymptotic Sampling Distributions (cont.)

Summary:

$$\sqrt{n}(M_{k,n} - \mu_k) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mu_{2k} - \mu_k^2 - 2k\mu_{k-1}\mu_{k+1} + k^2\mu_{k-1}^2\mu_2)$$

We could work out the asymptotic joint distribution of all these, but we spare you these details.

The $k = 2$ case is particularly simple. Recall $\mu_1 = 0$, $\mu_2 = \sigma^2$, and $M_{2,n} = V_n$, so the $k = 2$ case is

$$\sqrt{n}(V_n - \sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mu_4 - \sigma^4)$$

Asymptotic Sampling Distributions (cont.)

We will do one joint convergence in distribution result because we have already done all the work

$$\sqrt{n} \begin{pmatrix} \overline{X}_n - \mu \\ V_n - \sigma^2 \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} U \\ W \end{pmatrix}$$

or

$$\sqrt{n} \begin{pmatrix} \overline{X}_n - \mu \\ V_n - \sigma^2 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M})$$

where

$$\mathbf{M} = \begin{pmatrix} \mu_2 & \mu_3 \\ \mu_3 & \mu_4 - \mu_2^2 \end{pmatrix}$$

Asymptotic Sampling Distributions (cont.)

In contrast to the case where the data are exactly normally distributed, in general, $\overline{X}_n$ and V_n are not independent and are not even asymptotically uncorrelated unless the population third central moment is zero (as it would be for any symmetric population distribution but would not be for any skewed population distribution).

Moreover, in general, the asymptotic distribution of V_n is different from what one would get if a normal population distribution were assumed (homework problem).

Sampling Distribution of Order Statistics

Recall that $X_{(k)}$ is the k-th data value in sorted order. Its distribution function is

$$
\begin{aligned}
F_{X_{(k)}}(x) &= \Pr(X_{(k)} \leq x) \\
&= \Pr(\text{at least } k \text{ of the } X_i \text{ are} \leq x) \\
&= \sum_{j=k}^{n} \binom{n}{j} F(x)^j [1 - F(x)]^{n-j}
\end{aligned}
$$

Where

$$
F(x) = \Pr(X_i \leq x)
$$

Sampling Distribution of Order Statistics

If the data are continuous random variables having PDF $f = F'$, then the PDF of $X_{(k)}$ is given by

$$
\begin{aligned}
f_{X_{(k)}}(x) &= F'_{X_{(k)}}(x) \\
&= \frac{d}{dx} \sum_{j=k}^{n} \binom{n}{j} F(x)^j [1 - F(x)]^{n-j} \\
&= \sum_{j=k}^{n} \binom{n}{j} j F(x)^{j-1} f(x) [1 - F(x)]^{n-j} \\
&\qquad - \sum_{j=k}^{n-1} \binom{n}{j} F(x)^j (n-j) [1 - F(x)]^{n-j-1} f(x)
\end{aligned}
$$

Sampling Distribution of Order Statistics (cont.)

Rewrite the second term replacing j by $j - 1$ so the powers of $F(x)$ and $1 - F(x)$ match the first term

$$
\begin{aligned}
f_{X_{(k)}}(x) = \sum_{j=k}^{n} \binom{n}{j} j F(x)^{j-1} f(x) [1 - F(x)]^{n-j} \\
- \sum_{j=k+1}^{n} \binom{n}{j-1} F(x)^{j-1} (n - j + 1) [1 - F(x)]^{n-j} f(x) \\
= \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} [1 - F(x)]^{n-k} f(x)
\end{aligned}
$$

Sampling Distribution of Order Statistics (cont.)

If X_1, ..., X_n are IID from a continuous distribution having PDF f and DF F, then the PDF of the k-th order statistic is

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

and of course the domain is restricted to be the same as the domain of f.

Sampling Distribution of Order Statistics (cont.)

In particular, if X_1, ..., X_n are IID Unif$(0, 1)$, then the PDF of the k-th order statistic is

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}, \qquad 0 < x < 1$$

and this is the PDF of a Beta$(k, n-k+1)$ distribution.

Normal Approximation of the Beta Distribution

We cannot get a normal approximation directly from the CLT because there is no "addition rule" for the beta distribution (sum of IID beta does not have a brand name distribution).

Again we use the theorem: if X and Y are independent gamma distributions with the same rate parameter, then $X/(X + Y)$ is beta (5101 Deck 3, Slides 130–139, also used on slide 51 of this deck).

Normal Approximation of the Beta Distribution (cont.)

Suppose W is Beta(α_1, α_2) and both α_1 and α_2 are large. Then we can write

$$W = \frac{X}{X + Y}$$

where X and Y are independent gamma random variables with shape parameters α_1 and α_2, respectively, and the same rate parameter (say $\lambda = 1$).

Then we know that

$$X \approx \mathcal{N}(\alpha_1, \alpha_1)$$
$$Y \approx \mathcal{N}(\alpha_2, \alpha_2)$$

and X and Y are asymptotically independent (5101, Deck 6, Slide 85).

Normal Approximation of the Beta Distribution (cont.)

That is,

$$\begin{pmatrix} X \\ Y \end{pmatrix} \approx \mathcal{N}(\boldsymbol{\mu}, \mathbf{M})$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

$$\mathbf{M} = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix}$$

Normal Approximation of the Beta Distribution (cont.)

We now use the multivariate delta method to find the approximate normal distribution of W. (This is all a bit sloppy because we are using the "sloppy" version of the CLT. We could make it pedantically correct, but it would be messier.)

The transformation is $W = g(X, Y)$, where

$$
\begin{aligned}
g(x, y) &= \frac{x}{x + y} \\
\frac{\partial g(x, y)}{\partial x} &= \frac{y}{(x + y)^2} \\
\frac{\partial g(x, y)}{\partial y} &= -\frac{x}{(x + y)^2}
\end{aligned}
$$

Normal Approximation of the Beta Distribution (cont.)

The multivariate delta method says W is approximately normal with mean

$$g(\alpha_1, \alpha_2) = \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

and variance

$$\begin{aligned} \frac{1}{(\alpha_1 + \alpha_2)^4} \begin{pmatrix} \alpha_2 & -\alpha_1 \end{pmatrix} \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \begin{pmatrix} \alpha_2 \\ -\alpha_1 \end{pmatrix} &= \frac{\alpha_1 \alpha_2^2 + \alpha_1^2 \alpha_2}{(\alpha_1 + \alpha_2)^4} \\ &= \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^3} \end{aligned}$$

Normal Approximation of the Beta Distribution (cont.)

In summary

$$\text{Beta}(\alpha_1, \alpha_2) \approx \mathcal{N}\left(\frac{\alpha_1}{\alpha_1 + \alpha_2}, \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^3}\right)$$

when α_1 and α_2 are both large.

The parameters of the asymptotic normal distribution are no surprise, since the exact mean and variance of $\text{Beta}(\alpha_1, \alpha_2)$ are

$$E(W) = \frac{\alpha_1}{\alpha_1 + \alpha_2}$$
$$\text{var}(W) = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2 (\alpha_1 + \alpha_2 + 1)}$$

(brand name distributions handout) and the difference between $\alpha_1 + \alpha_2$ and $\alpha_1 + \alpha_2 + 1$ is negligible when α_1 and α_2 are large.

Sampling Distribution of Order Statistics (cont.)

Theorem: Suppose U_1, U_2, ... are IID Unif$(0, 1)$. Suppose

$$\sqrt{n}\left(\frac{k_n}{n} - p\right) \to 0, \qquad \text{as } n \to \infty$$

and suppose V_n denotes the k_n-th order statistic of U_1, ..., U_n, that is, for each n we sort U_1, ..., U_n and pick the k_n-th of these. Then

$$\sqrt{n}(V_n - p) \xrightarrow{\mathcal{D}} \mathcal{N}\bigl(0, p(1-p)\bigr)$$

or ("sloppy version")

$$V_n \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Sampling Distribution of Order Statistics (cont.)

Proof: The exact distribution of V_n is Beta$(k_n, n - k_n + 1)$. Hence

$$V_n \approx \mathcal{N}\left(\frac{k_n}{n+1}, \frac{k_n(n - k_n + 1)}{(n+1)^3}\right)$$

by the normal approximation for the beta distribution. Hence

$$\sqrt{n+1}\left(V_n - \frac{k_n}{n+1}\right) \approx \mathcal{N}\left(0, \frac{k_n(n - k_n + 1)}{(n+1)^2}\right)$$

Sampling Distribution of Order Statistics (cont.)

The right-hand side of the last display on the previous slide converges to $\mathcal{N}\big(0, p(1-p)\big)$ because

$$\begin{aligned} \frac{k_n}{n+1} &\to p \\ \frac{n - k_n + 1}{n+1} &\to 1 - p \end{aligned}$$

as $n \to \infty$. In summary,

$$\sqrt{n+1}\left(V_n - \frac{k_n}{n+1}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, p(1-p)\big)$$

Sampling Distribution of Order Statistics (cont.)

Now use Slutsky's theorem. Because of

$$\frac{n}{n+1} \to 1$$

$$\sqrt{n+1}\left(\frac{k_n}{n+1} - p\right) \to 0$$

as $n \to \infty$, we have

$$\sqrt{n+1}\left(V_n - \frac{k_n}{n+1}\right) = \sqrt{n}(V_n - p) + o_p(1)$$

and that finishes the proof.

Sampling Distribution of Order Statistics (cont.)

Now we use a result proved in 5101 homework problem 8-11. If U is a Unif$(0, 1)$ random variable, and G is the quantile function of another random variable X, then X and $G(U)$ have the same distribution.

The particular case of interest here X is a continuous random variable having PDF f which is nonzero on the support, which is an interval. If F denotes the DF corresponding to the quantile function G, then the restriction of F to the support is the inverse function of G. Hence by the inverse function theorem from calculus

$$\frac{dG(q)}{dq} = \frac{1}{\frac{dF(x)}{dx}} = \frac{1}{f(x)}$$

where $f = F'$ is the corresponding PDF, $x = G(q)$, and $q = F(x)$.

Sampling Distribution of Order Statistics (cont.)

Theorem: Suppose X_1, X_2, ... are IID from a continuous distribution having PDF f that is nonzero on its support, which is an interval. Let x_p denote the p-th quantile of this distribution. Suppose

$$\sqrt{n}\left(\frac{k_n}{n} - p\right) \to 0, \qquad \text{as } n \to \infty$$

and suppose V_n denotes the k_n-th order statistic of X_1, ..., X_n. Then

$$\sqrt{n}(V_n - x_p) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{p(1-p)}{f(x_p)^2}\right)$$

or ("sloppy version")

$$V_n \approx \mathcal{N}\left(x_p, \frac{p(1-p)}{nf(x_p)^2}\right)$$

Sampling Distribution of Order Statistics (cont.)

Proof: Use the univariate delta method on the transformation $X = G(U)$. Because functions of independent random variables are independent, we can write $X_i = G(U_i)$, where U_1, U_2, $\ldots$ are IID $\text{Unif}(0, 1)$. Then the univariate delta method says V_n is asymptotically normal with mean

$$G(p) = x_p$$

and variance

$$G'(p)^2 \cdot \frac{p(1-p)}{n} = \frac{1}{f(x_p)^2} \cdot \frac{p(1-p)}{n}$$

Sampling Distribution of the Sample Median

Theorem: Suppose X_1, X_2, ... are IID from a continuous distribution having PDF f that is nonzero on its support, which is an interval. Let m denote the median of this distribution, and suppose $\widetilde{X}_n$ denotes the sample median of X_1, ..., X_n. Then

$$\sqrt{n}(\widetilde{X}_n - m) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{4f(m)^2}\right)$$

or ("sloppy version")

$$\widetilde{X}_n \approx \mathcal{N}\left(m, \frac{1}{4nf(m)^2}\right)$$

Sampling Distribution of the Sample Median

Proof: If we only look at the n odd case, where the sample median is an order statistic, this follows from the previous theorem. The n even case is complicated by the conventional definition of the sample median as the average of the two middle order statistics. By the previous theorem these have the same asymptotic distribution (because $1/n \to 0$ as $n \to \infty$). Also they are ordered $X_{(n/2)} \le X_{(n/2+1)}$ always. Hence their asymptotic distribution, must also have this property. So assuming they do have an asymptotic joint distribution, it must be degenerate

$$X_{(n/2)} \approx X_{(n/2+1)} \approx \mathcal{N}\left(x_p, \frac{1}{4nf(m)^2}\right)$$

from which the theorem follows. We skip the details of proving that they are indeed jointly asymptotically normal.

Sampling Distribution of the Sample Median

What is the asymptotic distribution of the sample median of an IID sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution?

Since the normal distribution is symmetric, it mean and median are equal.

The normal PDF is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{(x-\mu)^2/2\sigma^2}$$

so

$$\widetilde{X}_n \approx \mathcal{N}\left(\mu, \frac{\pi\sigma^2}{2n}\right)$$