Paideia: Review of Statistics

Based mainly on HGL Probability Primer and Appendices B & C.

Random variables

- Variables that can take on one or more mutually exclusive **outcomes**.
- Which outcome will occur is unknown until it occurs and is **observed**.
- The **probability** of an outcome is the proportion of the time that the outcome occurs in repeated sampling.
- Random variables can be
    - **Discrete**, with a countable set of outcomes.
    - **Continuous**, with a continuum of possible outcomes.

Probability distributions

- **Discrete distributions**
    - List of all possible outcomes (sample space) and the probabilities that they will occur is **probability density function** of discrete r.v.
        - Notation: $f(x) \equiv \Pr[X = x]$.
    - Properties of density function:
        - $0 \leq f(x) \leq 1$
        - $\sum_{all\ x} f(x) = 1$
    - Cumulative distribution function of discrete r.v.
        - Notation: $F(x) \equiv \Pr[X \leq x]$.
    - Properties of cumulative distribution function
        - $F(x) = \sum_{\xi \leq x} f(\xi)$
        - $\Pr[X > x] = 1 - F(x)$
- **Continuous distributions**
    - With infinite number of outcomes in sample space, probability of any individual outcome is zero.
    - **Probability density function** is continuous function over the sample space whose value represents the relative likelihood of that outcome occurring (but not the probability, see above).
        - Probability of any range of outcomes is the definite integral of the density function over that interval: $\Pr[x_1 < X \leq x_2] = \int_{x_1}^{x_2} f(x)\,dx$.

- Integral of density function over entire sample space must equal one: $\int_{-\infty}^{\infty} f(X)\,dX = 1$.
- We often denote the density function is $f(\cdot)$.

o **Cumulative distribution function** is the integral of the density function: $F(x) = \Pr[X < x] = \int_{-\infty}^{x} f(X)\,dX = \int f(x)\,dx$.

- $\Pr[x_1 < X \le x_2] = \int_{x_1}^{x_2} f(X)\,dX = \int_{x_1}^{x_2} dF(X) = F(x_2) - F(x_1)$.
- This also implies that $f(x) = F'(x)$.

Multivariate distributions

In economics we are almost always interested in the joint variation in two or more variables. We will talk about bivariate distributions here, but the results easily generalize to more variables.

- **Joint distributions**
  - For discrete distributions: $f(x,y) \equiv \Pr[X = x, Y = y]$.
  - For continuous distributions: $\Pr[x_1 < X \le x_2, y_1 < Y \le y_2] = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(X,Y)\,dY\,dX$.
- **Marginal distributions**
  - Marginal distribution is the univariate distribution of one of the variables in a joint distribution.
  - Discrete: $f_X(x) = \Pr[X = x] = \sum_{all\ y} f(x,y). \in$
    - See example in P.4 on page 23.
  - Continuous: $f_X(x) = \int_{y=-\infty}^{\infty} f(x,y)\,dy$.
- **Conditional distributions**
  - Very important for econometrics.
  - Distribution of *y* given that *x* takes on a particular value.
    - Application: what is the distribution of Reed GPAs conditional on a student having perfect SAT scores (or any other particular value)?
  - Discrete: see example in Table P.5 on page 23.
  - Continuous marginal, joint, and conditional distributions are related by $f(y \mid x) = \dfrac{f(x,y)}{f_X(x)}$, or $f(x,y) = f(y \mid x) f_X(x)$.
  - Conditional distribution as a two-dimensional slice at a particular value of *x* from the three-dimensional joint probability distribution.
- **Independence**

- Two variables are independent if these equivalent conditions are satisfied for all x and y
  - $f(x \mid y) = f_X(x)$
  - $f(x, y) = f_X(x) \cdot f_Y(y)$

Expected values and moments

- **Expected value** is the average outcome expected over infinite draws from the distribution.
  - $E(X)$ is the notation for expected value of a random variable X.
  - $E(X)$ is also called the **population mean** of the distribution or the mean of X and is often denoted by μ_X.
  - For discrete random variables:
    - The mean X is $E(X) = \sum_{all\ x} x \cdot \Pr[X = x]$.
  - For continuous random variables:
    - The expected value of X is $\mu_X \equiv E(X) = \int_{-\infty}^{\infty} X\, f(X)\, dX = \int_{-\infty}^{\infty} X\, dF(X)$.
  - **Conditional expectation**: $E(Y \mid X)$ is the mean of the conditional distribution:

    $$E(Y \mid X = x) \equiv \int_{y=-\infty}^{\infty} y \cdot f(y \mid X = x)\, dy.$$

    - Note that x is given, so we don't integrate over X.
    - This is the mean of the conditional distribution. (Think of slicing the multivariate distribution at one X value.)
  - **Law of iterated expectations**:

    $$E(Y) = E_X\left[E(Y \mid X)\right] = \int_{x=-\infty}^{\infty} \left[\int_{y=-\infty}^{\infty} y \cdot f(y \mid X = x)\, dy \right] f_X(x)\, dx.$$

    - Note that the inner expectation will generally be a function of X and the outer expectation is over X.
  - Properties of expected values (where a is a constant and X and Y are random variables):
    - $E(aX) = aE(X)$,
$E(X + Y) = E(X) + E(Y)$.
    - Note that it is *not* generally true that $E(XY) = E(X)E(Y)$.
- We often characterize distributions by summary measures called **moments**.
  - The nth absolute moment of X is $E(X^n) = \int_{-\infty}^{\infty} X^n dF(X)$.
    - The mean is the first absolute moment of X.

- When thinking about higher-order moments, it is usually more convenient to work with moments around the mean, or central moments, rather than absolute moments.
  - The *n*th central moment of *X* is $E\left(\left[X-\mu_X\right]^n\right)=\int_{-\infty}^{\infty}\left[X-\mu_X\right]^n dF(X)$.
- The second central moment of a random variable is its **variance**:
  - $\sigma_X^2 \equiv \text{var}(X) \equiv E\left[\left(X-\mu_X\right)^2\right]=E\left(X^2\right)-\mu_X^2$.
  - Because the units of the variance are the square of the units of *X*, we often find its square root, the **standard deviation**, more useful. Since the variance is σ^2, the standard deviation is just σ.
  - Variance of a conditional distribution is called the **conditional variance**.
- **Covariance and correlation**
  - Covariance: $\sigma_{XY} \equiv \text{cov}(X,Y) \equiv E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]$.
    - $\text{cov}(X,X)=\text{var}(X)$, so covariance is just a straightforward generalization of variance.
    - Like variance, covariance is in awkward units: product of units of *X* and units of *Y*.
  - Correlation coefficient is unit free:

    $$\rho_{XY} \equiv \text{corr}(X,Y) \equiv \frac{\sigma_{XY}}{\sigma_X \sigma_Y}=\frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}.$$

    - $-1 \le \rho_{XY} \le 1$.
    - $\rho_{XY} = 0$ means variables are uncorrelated and cov = 0.
    - Independent random variables are always uncorrelated (but the converse is not always true).
    - $E(Y \mid X)=E(Y) \Rightarrow \text{cov}(X,Y)=0$, but converse is not always true.
  - Properties of variances of random variables:
    - $$\begin{aligned}&\text{var}(aX)=a^2\,\text{var}(X),\\ &\text{var}(X+Y)=\text{var}(X)+\text{var}(Y)+2\,\text{cov}(X,Y),\end{aligned}$$
    - Variance of a constant is always zero.
- Third central moment (divided by the variance) is the coefficient of **skewness**.
  - A value of zero for skewness indicates a symmetric matrix.
  - Positive skewness reflects a long right tail; negative skewness a long left tail.
- Fourth central moment (again, divided by the variance) is the **kurtosis** of the distribution.
  - High kurtosis means heavy tail on the distribution.

- (High-kurtosis distributions have become very important in mathematical finance.)
- Normal distribution is the neutral standard with kurtosis of 3.

Useful probability distributions

- **Bernoulli distribution**
    - $X = 1$ with probability p and 0 otherwise
    - $f(x) = p^x (1-p)^{1-x}$
    - $E(X) = p$
    - $\text{var}(X) = p(1-p)$
- **Binomial distribution**
    - Sum of N independent Bernoulli variables
    - $f(x) = \binom{N}{x} p^x (1-p)^{N-x}, \quad \binom{N}{x} \equiv \frac{N!}{x!(N-x)!}$
    - $E(x) = Np$
    - $\text{var}(x) = Np(1-p)$
- **Poisson distribution**
    - Models number of Bernoulli occurrences in given span of time
    - $f(x) = \frac{e^{-\mu}\mu^x}{x!}$, where μ is expected number of occurrences per unit of time
    - $E(x) = \text{var}(x) = \mu$
- **Uniform distribution**
    - Equal probability of all outcomes over range a to b
    - $f(x) = \frac{1}{b-a}$
    - $E(x) = \frac{a+b}{2}$
    - $\text{var}(x) = \frac{(b-a)^2}{12}$
- **Normal (Gaussian) distribution**
    - $\phi(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2}$, (no closed form for Φ, the cumulative distribution function)
        - Normal has two parameters: mean and variance
        - $N(\mu_X, \sigma_X^2)$ is standard notation for the normal distribution with mean μ and variance σ^2

- **Standard normal** has mean 0 and variance 1.
  - Can always convert normal X to standard normal Z:
    - If $X \sim N(\mu, \sigma^2)$, then $Z \equiv \frac{X-\mu}{\sigma} \sim N(0, 1)$.
- Normal distribution is closed on addition, subtraction, and scalar multiplication.
- **Multivariate normal**:
  - Bivariate normal distribution is fully characterized by five parameters: two means, two variances, and one covariance (or correlation)
  - For jointly normal variables (but not in general case) $\rho = 0$ implies independence.
- $aX + bY \sim N\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}\right)$ if X and Y are jointly normal and a and b are constants.

- **Chi-square distribution**
  - The sum of the squares of M independent, standard normal random variables is distributed as a χ^2 with M degrees of freedom.
  - The χ^2 distribution has one parameter M, the number of degrees of freedom, and has mean equal to M and variance $2M$.
  - The χ^2 distribution is only defined for positive values and is skewed to the right.
  - The χ^2 distribution converges to the normal as $M \to \infty$.
- **Student *t* distribution**
  - Named for its discoverer (which went by the pseudonym "Student," the t distribution with M degrees of freedom is the distribution that is followed by $\frac{Z}{\sqrt{W/M}}$, where $Z \sim N(0, 1)$, $W \sim \chi_M^2$, and Z and W are independent.
  - The t distribution is symmetric but with larger kurtosis than the normal.
  - It converges to the standard normal as $M \to \infty$.
- ***F* distribution**
  - If $W \sim \chi_M^2$ and $V \sim \chi_N^2$, then $\frac{W/M}{V/N} \sim F_{M,N}$.
  - The F distribution has two parameters, the numerator and denominator degrees of freedom.
  - The F distribution is defined only over positive values. Its mean is one.
  - The $F_{M,N}$ distribution converges to χ_M^2 as $N \to \infty$.

Populations, data-generating processes, and samples

- From where do our data come?
  - Often, we think of a cross-sectional **population** from which we have drawn a sample.

- In time series, we usually think of an infinite **data-generating process** of which our sample is a finite realization.

- **Estimation of population parameters**
  - We usually make some assumptions about the general nature of the population distribution of a random variable (e.g., the population distribution is normal) that leaves some unknown parameters to be estimated (the mean and variance).
  - Statistics involves methods of estimating unknown population (or DGP) parameters by calculating **test statistics** based on a **random sample**.
- **Random sampling**
  - Does each element in the population have an equal probability of being sampled (conditional on the others having been sampled)?
    - What is the relevant population?
    - In phone survey, the population consists of people who have and answer phones (or perhaps people in the phone book who have and answer).
  - The **observations**—draws from the population (realizations of the data-generating process)—comprise our **sample**.
    - We usually denote the number of observations in the sample by *n*.
  - The observations are random variables, as are functions of them.
  - If each observation drawn to be in the sample follows the same distribution and is independent of all other draws, then the sample is **independently and identically distributed (IID)**.
- **Sample moments**
  - **Sample mean**
    - $\bar{y} \equiv \frac{1}{N}\sum_{i=1}^{N} y_i$ = average of the sample values.
      - HGL use $\bar{Y}$ for the variable and $\bar{y}$ for the calculated value.
    - $E(Y_i) = \mu_Y$ if sample is drawn randomly, so $E(\bar{Y}) = \frac{1}{N}\sum_{i=1}^{n} E(Y_i) = \mu_Y$.
      - The sample mean $\bar{Y}$ is an **unbiased** estimator of the population mean.
    - Since all observations are assumed to have the same variance, $\text{var}(\bar{Y}) = \sum_{i=1}^{N}\left[\frac{1}{N^2}\text{var}(Y_i)\right] + \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N}\frac{1}{N^2}\text{cov}(Y_i, Y_j)$. (Note that we don't need the 2 in front of the double sum because each *i, j* pair is picked up twice in the summation.)
    - If the sample is IID, then cov(Y_i, Y_j) = 0 for $i \neq j$, so $\text{var}(\bar{Y}) = N \cdot \left[\frac{1}{N^2}\text{var}(Y_i)\right] = \frac{1}{N}\sigma_Y^2$.

            - This is interesting and helpful because the variance of the sample mean $\to 0$ as the sample gets large.
        - Since $\bar{Y}$ is a linear function of the sample observations, it is normally distributed if the population is normal.
            - If $Y_i \sim N\left(\mu_Y, \sigma_Y^2\right)$, then $\bar{Y} \sim N\left(\mu_Y, \frac{1}{N}\sigma_Y^2\right)$.
- **Sample variance**
    - $\hat{\sigma}_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$.
        - We divide by $N-1$ rather than N because of "degrees of freedom."
        - Only $N-1$ of the terms being summed are independent because the (non-squared) terms in parentheses add to zero.
- **Estimated variance of sample mean**
    - $\text{var}\left(\bar{Y}\right) = \frac{\sigma_Y^2}{N}$, but we don't know σ_Y^2
    - To estimate variance of sample mean, substitute estimated variance of Y: $\widehat{\text{var}\left(\bar{Y}\right)} = \frac{\hat{\sigma}_Y^2}{N}$
    - **Standard error** is square root of estimated variance
- **Sample covariance and correlation coefficients**
    - Sample covariance is $\hat{\sigma}_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$.
    - Sample correlation is $r_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$.
- **Sample skewness and kurtosis**
    - $S = \frac{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^3 / N}{\left[\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 / N\right]^{3/2}}$
    - $K = \frac{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^4 / N}{\left[\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 / N\right]^{2}}$
    - We can use the sample skewness and kurtosis to test whether a sample comes from a normal population, since the skewness and kurtosis of a standard normal variable are 0 and 3.

Asymptotic distributions

If we know (or assume) the exact population distribution, then we can often calculate the exact distribution of our estimators based on the sample. We often assume (rightly or wrongly) that our data com from normal distributions. More commonly, we may not know that the population is normal (or anything else). In these cases, we may still be able to know something about how the sample estimators behave **asymptotically**, as the sample size gets large.

- **Convergence in distribution**
    - If the probability distribution of a random variable becomes arbitrarily close to some limiting probability distribution as $N \to \infty$, then we say that the variable converges in distribution to the limiting distribution.
        - We shall formalize this mathematically during the course.
- **Convergence in probability**
    - If the probability that a random variable differs from a constant a by more than an arbitrarily small amount δ approaches zero as $N \to \infty$, then we say that the random variable converges in probability to a, or the probability limit of the random variable is a.
    - A statistic that converges in probability to the parameter it is intended to estimate is said to be **consistent**.
- The **law of large numbers** assures us that $\text{plim}\,\bar{X} = \mu_X$ for any IID sample from a population with constant mean and finite variance.
- The **central limit theorem** assures us that if we sample IID from a population with constant mean and finite variance, $\bar{Y} \xrightarrow{d} N\left(\mu_Y, \frac{1}{N}\sigma_Y^2\right)$.
    - This means that even if the population is not normally distributed, the sample mean will (under quite general conditions) converge to a normal distribution as the sample size gets large.

Estimators

- An **estimator** is a function of the sample observations that is intended to provide information about a population parameter.
    - An estimator is a formula or function that takes on various values depending on the sample that is drawn.
    - The value or the estimator calculated based on a specific sample is called the **estimate**.
    - Estimators are random variables because they depend on the sample values, which are random variables.

- We explored the sample mean as an estimator of the population mean.
  - We characterized the distribution of the sample mean:
    - $E(\bar{Y}) = \mu_Y$,
    - $\text{var}(\bar{Y}) = \frac{1}{N}\sigma_Y^2$,
    - $\bar{Y}$ is asymptotically normal.
- **Unbiasedness** means that the expected value of the estimator equals the parameter it is intended to estimate. Let $\hat{\theta}$ be an estimator for a parameter θ.
  - The bias in $\hat{\theta}$ is $E(\hat{\theta}) - \theta$.
  - If the bias = 0, then the estimator is unbiased.
    - Because $E(\bar{Y}) = \mu_Y$, it is an unbiased estimator of μ_X.
- An estimator $\hat{\theta}$ is **consistent** if plim $\hat{\theta}$ = θ.
  - Unbiasedness is neither necessary nor sufficient for consistency.
    - Bias can go to zero as sample gets large, so biased estimator can be consistent.
      - For example, $\frac{1}{N-1}\sum_{i=1}^{N} Y_i$ is a biased, but consistent estimator for μ_Y.
    - Variance of unbiased estimator may not go to zero.
      - For example, Y_1 is an unbiased but inconsistent estimator for μ_Y.
  - Because plim $\bar{Y}$ = μ_Y, the sample mean is a consistent estimator of the population mean.
  - Sample variance and covariance are unbiased and consistent estimators of the population values as well.
- An estimator is **efficient** if it has minimum variance among all unbiased estimators.
- An estimator is **asymptotically efficient** if its variance goes to zero at least as fast as all consistent estimators.

Hypothesis tests

- A formal **hypothesis test** involves specifying a **null hypothesis** (that we usually wish to disprove) and an **alternative hypothesis** that holds if the null hypothesis is false.
  - For example, we might use the null and alternative hypotheses $H_0 : \mu_Y = 3, H_1 : \mu_Y \neq 3$.
  - The alternative hypothesis can be one-sided or two-sided.
- The test either **rejects** or **fails to reject** the null hypothesis based on our sample.
  - If it is sufficiently unlikely that a sample that is this deviant from the null hypothesis would occur randomly, we reject the null.

- **Type I and Type II error**:

		Actual null hypothesis is:	
		True	False
Test result is:	Accept (fail to reject) null	Correct conclusion	**Type II error**
	Reject null	**Type I error**	Correct Conclusion

- In a hypothesis test, we choose a **significance level** to be the probability of Type I error.
  - 5% is conventional, but 10%, 1%, and even 0.1% are sometimes used.
- The ***p*-value** of a test statistic based on a sample is the probability of drawing a sample whose test statistic differs at least as much from the null hypothesis as the one we have drawn.
- We reject the null hypothesis if the *p*-value is less than our significance level.
- For sample mean, we know that $\frac{\bar{Y}-\mu_Y}{\sigma_Y / \sqrt{N}} \sim N(0,1)$, at least asymptotically.
  - If we know σ_Y, then we can calculate this test statistic under the null hypothesis value c of μ_Y.
    - $Z = \frac{\bar{Y}-c}{\sigma_Y / \sqrt{N}} \sim N(0,1)$,
  - How far out in the tails of the normal distribution does the test statistic fall?
  - The normal distribution has 2.5% of its density in the (two) tails beyond the values ±1.96. So we reject the null hypothesis at a significance level of 5% if the absolute value of the test statistic exceeds 1.96.
- If we don't know σ_Y, we must approximate it by the sample variance.
  - Based on our estimate $\hat{\sigma}_Y^2$ of σ_Y^2, we can calculate an estimate (called the **standard error**) of the standard deviation of $\bar{Y}$ as $SE(\bar{Y}) = \frac{1}{\sqrt{N}}\hat{\sigma}_Y$.
  - Because $\hat{\sigma}_Y^2$ is proportional to a χ^2 variable with $N - 1$ degrees of freedom and is independent of the normally distributed numerator, $\frac{\bar{Y}-\mu_Y}{SE(\bar{Y})}$ follows a t distribution with $N - 1$ degrees of freedom.
  - Under the null hypothesis $\mu_Y = c$, the test statistic $t = \frac{\bar{Y}-c}{SE(\bar{Y})} \sim t_{N-1}$, so we reject the null hypothesis if the test statistic falls in the upper or lower 2.5% tails of the t distribution

One-tailed tests

- Sometimes we are only interested in one-sided alternative hypotheses: H_1: $\mu_Y < c$.

- In this case we pull all 5% of the rejection probability from the lower tail of the distribution and fail to reject the null for anything in the upper half of the distribution

- Similarly, if H_1: $\mu_Y > c$, then we reject only the upper tail of the distribution.

p-values

- The *p*-value of a test statistic is the smallest level of significance at which we can reject the null hypothesis.
- In a two-tailed test of H_0: $\Theta = c$, if the test statistic is $\hat{\theta}$ and $\Pr\left[|\Theta - c| > |\hat{\theta} - c|\right] = p_0$, then p_0 is the *p*-value for the statistic.
- In a test with test statistic t, the *p*-value is:
  - the probability to the right of t if H_1: $\mu > c$
  - the probability to the left of t if H_1: $\mu > c$
  - the probability to the right of $|t|$ and to the left of $-|t|$ if H_1: $\mu \neq c$
- *p*-values make it easy to determine the result of the test: reject if $p < \alpha$

Confidence intervals

- A confidence interval is a region that, $(1 - \alpha)$ share of the time, will contain the true population parameter.
- For example, if $Y \sim N(\mu_Y, \sigma_Y^2)$, then $\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{N}\right)$, and $z = \frac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{N}} \sim N(0,1)$.
  - From the tables of the normal distribution, we know that $\Pr\left[-1.96 < z < 1.96\right] = 0.05$.
  - Thus,

$$
\begin{aligned}
0.95 &= \Pr\left[-1.96 < \frac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{N}} < 1.96\right] \\
&= \Pr\left[-1.96 \frac{\sigma_Y}{\sqrt{N}} < \bar{Y} - \mu_Y < 1.96 \frac{\sigma_Y}{\sqrt{N}}\right] \\
&= \Pr\left[-1.96 \frac{\sigma_Y}{\sqrt{N}} - \bar{Y} < -\mu_Y < 1.96 \frac{\sigma_Y}{\sqrt{N}} - \bar{Y}\right] \\
&= \Pr\left[\bar{Y} - 1.96 \frac{\sigma_Y}{\sqrt{N}} < \mu_Y < \bar{Y} + 1.96 \frac{\sigma_Y}{\sqrt{N}}\right].
\end{aligned}
$$

  - This is our 95% confidence interval for the population mean.
  - Note the symmetry between the hypothesis test and the confidence interval
    - We reject the null hypothesis that the parameter equals a at the α level of significance if a does not fall within the $(1 - \alpha)$ confidence interval.