Paideia: Review of Statistics

Based mainly on S&W Ch. 2 & 3, but integrating some material from GHJ.

Random variables

- Variables that can take on one or more mutually exclusive **outcomes**.
- 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 finite set of outcomes.
    - **Continuous**, with a continuum of possible outcomes.
- Sample space is the set of all possible outcomes and an event is a subset of the sample space.

Probability distributions

- **Discrete distributions**
    - List of all possible outcomes (sample space) and the probabilities that they will occur
    - Probability density function of discrete r.v.
    - Probability of events = sum of probabilities of mutually exclusive outcomes in the event.
    - Cumulative probability function of discrete r.v.
    - Bernoulli r.v.
        - Outcomes are 0, 1
        - Pr[X = 1] = p
- **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: $\Pr[x_1 < X \le x_2] = \int_{x_1}^{x_2} f(X)dX.$
        - 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 $f(\cdot)$.

- **Cumulative distribution function** is the integral of the density function:

  $F(x) = \Pr[X < x] = \int_{-\infty}^{x} f(X)dX = \int f(x)dx.$

  - $\Pr[x_1 < X \le x_2] = \int_{x_1}^{x_2} f(X)dX = \int_{x_1}^{x_2} dF(X) = F(x_2) - F(x_1).$

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 **mean** of the distribution or the mean of X and is often denoted by μ_X.
  - We can also take expected values of functions of random variables or functions of random variables based on the random variable's distribution.
  - For discrete random variables:
    - $E(g(X)) = \sum_{x \in S} g(x) \cdot \Pr[X = x]$, where S is the sample space.
    - The mean of the random variable itself is $E(X) = \sum_{x \in S} x \cdot \Pr[X = x].$
  - For continuous random variables:
    - $E(g(X)) \equiv \int_{-\infty}^{\infty} g(X) f(X) dX.$
    - This is sometimes written as $E(g(X)) \equiv \int_{-\infty}^{\infty} g(X) dF(X)$ recognizing that $\frac{dF(X)}{dX} = f(X)$, so $f(X)dX = dF(X).$
    - So, the expected value of X itself is just

      $\mu_X \equiv E(X) = \int_{-\infty}^{\infty} X f(X) dX = \int_{-\infty}^{\infty} X \, dF(X).$
  - 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 nth 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**:
  - $\sigma_X^2 \equiv \text{var}(X) \equiv E\left[\left(X-\mu_X\right)^2\right]$.
  - 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.
  - Properties of variances of random variables:
    - $\text{var}(aX)=a^2\,\text{var}(X)$, $\text{var}(X+Y)=\text{var}(X)+\text{var}(Y)+2\,\text{cov}(X,Y)$, where cov will be defined later.
    - 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.

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: $f(x)=\Pr[X=x]=\sum_{y\in S_Y} f(x,y)$.
    - See example in Table 2.2 on page 30.

- Continuous: $f(x) = \int_{-\infty}^{\infty} f(x,Y)dY$.

- **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 2.3 on page 31.
  - Continuous marginal, joint, and conditional distributions are related by

    $f(y \mid x) = \frac{f(x,y)}{f(x)}$, or $f(x,y) = f(y \mid x) f(x)$.
  - Conditional distribution as a two-dimensional slice at a particular value of X from the three-dimensional joint probability distribution.
  - **Conditional expectation**: $E(Y|X)$ is the mean of the conditional distribution:

    $E(Y \mid X = x) \equiv \int_{-\infty}^{\infty} Y \cdot f(Y \mid X = x)dY$.
    - 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**: $E(Y) = E\left[E(Y \mid X)\right]$.
    - Note that the inner expectation will generally be a function of X and the outer expectation is over X.
  - **Conditional variance**: $\text{var}(Y \mid X) \equiv E\left[\left(Y - E(Y \mid X)\right)^2 \mid X\right]$.
    - This is just the variance of the conditional distribution.
- **Statistical independence**
  - Two variables are statistically independent if $f(Y) = f(Y \mid X), \forall X,$ which leads to $f(X,Y) = f(X) f(Y)$.
- **Covariance and correlation**
  - Covariance: $\sigma_{XY} \equiv \text{cov}(X,Y) \equiv E\left[(X - \mu_X)(Y - \mu_Y)\right]$.
    - $\text{cov}(X,X) = \text{var}(X)$, 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.

Useful probability distributions

- **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) ρ = 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 both mean and variance equal to *m*.
  - 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 $Z \sim N(0, 1)$, $W \sim \chi^2_m$, 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 $W \sim \chi^2_m$ and $V \sim \chi^2_n$, 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^2_m 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.
- **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**
    - $\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$ = average of the sample values.
    - $E(X_i) = \mu_X$ if sample is drawn randomly, so $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu_X$.

- The sample mean $\bar{X}$ is an **unbiased** estimator of the population mean.

- Since all observations are assumed to have the same variance, $\text{var}\left(\bar{X}\right) = \sum_{i=1}^{n}\left[\frac{1}{n^2}\text{var}\left(X_i\right)\right] + \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\frac{1}{n^2}\text{cov}\left(X_i, X_j\right)$. (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 $\text{cov}(X_i, X_j) = 0$ for $i \neq j$, so $\text{var}\left(\bar{X}\right) = n \cdot \left[\frac{1}{n^2}\text{var}\left(X_i\right)\right] = \frac{1}{n}\sigma_X^2$.
  - This is interesting and helpful because the variance of the sample mean $\to 0$ as the sample gets large.
- Since $\bar{X}$ is a linear function of the sample observations, it is normally distributed if the population is normal.
  - If $X_i \sim N\left(\mu_X, \sigma_X^2\right)$, then $\bar{X} \sim N\left(\mu_X, \frac{1}{n}\sigma_X^2\right)$.

- **Sample variance**
  - $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$.
    - 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.
- **Sample covariance and correlation coefficients**
  - Sample covariance is $s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$.
  - Sample correlation is $r_{XY} = \frac{s_{XY}}{s_X s_Y}$.

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 $\operatorname{plim} \bar{X} = \mu_X$ 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, $\bar{X} \xrightarrow{d} N\left(\mu_X, \frac{1}{n}\sigma_X^2\right)$.
  - 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:
      - $E\left(\bar{X}\right) = \mu_X$,
      - $\operatorname{var}\left(\bar{X}\right) = \frac{1}{n}\sigma_X^2$,
      - $\bar{X}$ 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 $E\left(\hat{\theta}\right) - \theta$.

- If the bias = 0, then the estimator is unbiased.
  - Because $E(\bar{X}) = \mu_X$, it is an unbiased estimator of μ_X.

- An estimator $\hat{\theta}$ is **consistent** if plim $\hat{\theta} = \theta$.
  - 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, $\frac{1}{n-1}\sum_{i=1}^{n} X_i$ is a biased, but consistent estimator for μ_X.
    - Variance of unbiased estimator may not go to zero.
      - For example, X_1 is an unbiased but inconsistent estimator for μ_X.
  - Because plim $\bar{X} = \mu_X$, 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 $H_0 : \mu_X = 3, H_1 : \mu_X \neq 3$.
  - 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 $\frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}} \sim N(0,1),$ at least asymptotically.
    - If we know σ_X, then we can calculate this test statistic under the null hypothesis value of μ_X.
    - 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_X, we must approximate it by the sample variance.
    - Based on our estimate s_X^2 of σ_X^2, we can calculate an estimate (called the **standard error**) of the standard deviation of $\bar{X}$ as $SE(\bar{X}) = \frac{1}{\sqrt{n}} s_X.$
    - Because s_X^2 is proportional to a χ^2 variable with $n - 1$ degrees of freedom and is independent of the normally distributed numerator, $\frac{\bar{X} - \mu_X}{SE(\bar{X})}$ follows a *t* distribution with $n - 1$ degrees of freedom.

Confidence intervals

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

- Thus,

$$0.95 = \Pr\left[-1.96 < \frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}} < 1.96 \right]$$

$$= \Pr\left[-1.96 \frac{\sigma_X}{\sqrt{n}} < \bar{X} - \mu_X < 1.96 \frac{\sigma_X}{\sqrt{n}} \right]$$

$$= \Pr\left[-1.96 \frac{\sigma_X}{\sqrt{n}} - \bar{X} < -\mu_X < 1.96 \frac{\sigma_X}{\sqrt{n}} - \bar{X} \right]$$

$$= \Pr\left[\bar{X} - 1.96 \frac{\sigma_X}{\sqrt{n}} < \mu_X < \bar{X} + 1.96 \frac{\sigma_X}{\sqrt{n}} \right].$$

- 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.