Review of Elementary Statistics

Economics 143 Jinyong Hahn

Comments by Michael Powell

Random Variables

- Variables that takes on alternative values, each with a probability less than or equal to 1
- Probability distribution associates all possible outcomes with the probability that each will occur
- Continuous random variable: Any value on real line
- Discrete random varaible: Finite number of real values

Expected Value

- X: Discrete random variable
- $x_1, \ldots, x_N$: N possible outcomes
- $p_1, \ldots, p_N$: Associated probabilities
- Expected value of the mean is then defined by

$$E[X] = \mu_X = p_1 x_1 + \cdots + p_N x_N = \sum_{i=1}^{N} p_i x_i$$

Variance

- Variance provides a measure of the dispersion of random variable around the mean
- Weighted average of the squares of the deviations of outcomes from its expected value

$$Var[X] = \sigma_X^2 = \sum_{i=1}^{N} p_i x_i^2 - E[X]^2$$

Comment: There is an alternative definition of variance that is common in most textbooks:

$$Var[X] = \sum_{i=1}^{N} (x_i - E[X])^2 p_i$$

This is equivalent to the above definition:

$$\begin{aligned}\sum_{i=1}^{N}\left(x_i - E\left[X\right]\right)^2 p_i &= \sum_{i=1}^{N}\left(x_i^2 - 2x_iE\left[X\right] + \left(E\left[X\right]\right)^2\right)p_i \\ &= \sum_{i=1}^{N}x_i^2p_i - 2E\left[X\right]\sum_{i=1}^{N}x_ip_i + \left(E\left[X\right]\right)^2\sum_{i=1}^{N}p_i \\ &= E\left[X^2\right] - 2E\left[X\right]E\left[X\right] + \left(E\left[X\right]\right)^2 \\ &= E\left[X^2\right] - \left(E\left[X\right]\right)^2\end{aligned}$$

- Variance is in itself an expectation
- The (positive) square root of the variance is called the standard deviation, and written as σ_X

Useful Results

Proposition: $E\left[aX+b\right] = aE\left[X\right] + b$

Proof:

$$\begin{aligned}E\left[aX+b\right] &= \sum_{i=1}^{N}\left(ax_i + b\right)p_i \\ &= a\sum_{i=1}^{N}x_ip_i + b\sum_{i=1}^{N}p_i \\ &= aE\left[X\right] + b \;\text{ since } \sum_{i=1}^{N}p_i = 1\end{aligned}$$

Proposition: $E\left[\left(aX\right)^2\right] = a^2E\left[X^2\right]$

Proof:

$$\begin{aligned}E\left[\left(aX\right)^2\right] &= \sum_{i=1}^{N}\left(ax_i\right)^2p_i \\ &= \sum_{i=1}^{N}a^2x_i^2p_i \\ &= a^2\sum_{i=1}^{N}x_i^2p_i \\ &= a^2E\left[X^2\right]\end{aligned}$$

Proposition: $Var\left[aX+b\right] = a^2Var\left[X\right]$

Proof:

$$\begin{aligned}
Var\left[aX+b\right] &= \sum_{i=1}^{N}\left(\left(ax_i+b\right)-E\left[aX+b\right]\right)^2 p_i \\
&= \sum_{i=1}^{N}\left(ax_i+b-aE\left[X\right]-b\right)^2 p_i \text{ by above proposition} \\
&= \sum_{i=1}^{N}\left(a\left(x_i-E\left[X\right]\right)\right)^2 p_i \\
&= a^2\sum_{i=1}^{N}\left(x_i-E\left[X\right]\right)^2 p_i \\
&= a^2 Var\left[X\right]
\end{aligned}$$

Two Random Variables

- (X,Y): Discrete random variables
- Joint distribution can be described by a list of all possible outcomes $(x_i, y_j) \quad i=1,\ldots,M; \quad j=1,\ldots,N$, and associated probabilities p_{ij}
- Example: X is the family income. $Y=1$ if the head of ahousehold has a college education, and $Y=0$ otherwise.

	$Y=1$	$Y=0$
$X=5000$	0	1/4
$X=10,000$	1/8	1/8
$X=15,000$	1/3	1/6

Covariance

$$\begin{aligned}
Cov\left(X,Y\right) &= E\left[\left(X-E\left[X\right]\right)\left(Y-E\left[Y\right]\right)\right] \\
&= E\left[XY\right]-E\left[X\right]E\left[Y\right] \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N} p_{ij}x_iy_j - E\left[X\right]E\left[Y\right]
\end{aligned}$$

Proof:

$$\begin{aligned}
Cov\left(X,Y\right) &= E\left[\left(X-E\left[X\right]\right)\left(Y-E\left[Y\right]\right)\right] \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N}\left(x_i-E\left[X\right]\right)\left(y_j-E\left[Y\right]\right)p_{ij} \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N}\left(x_iy_j-y_jE\left[X\right]-x_iE\left[Y\right]+E\left[X\right]E\left[Y\right]\right)p_{ij} \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N}x_iy_jp_{ij}-E\left[X\right]\sum_{j=1}^{M}\sum_{i=1}^{N}y_jp_{ij}-E\left[Y\right]\sum_{j=1}^{M}\sum_{i=1}^{N}x_ip_{ij}+E\left[X\right]E\left[Y\right]\sum_{j=1}^{M}\sum_{i=1}^{N}p_{ij} \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N}x_iy_jp_{ij}-E\left[X\right]\sum_{j=1}^{M}y_j\sum_{i=1}^{N}p_{ij}-E\left[Y\right]\sum_{i=1}^{N}x_i\sum_{j=1}^{M}p_{ij}+E\left[X\right]E\left[Y\right]\sum_{j=1}^{M}p_j^Y \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N}x_iy_jp_{ij}-E\left[X\right]\sum_{j=1}^{M}y_jp_j^Y-E\left[Y\right]\sum_{i=1}^{N}x_ip_i^X+E\left[X\right]E\left[Y\right]\sum_{j=1}^{M}p_j^Y \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N}x_iy_jp_{ij}-E\left[X\right]E\left[Y\right]-E\left[Y\right]E\left[X\right]+E\left[X\right]E\left[Y\right] \\
&= E\left[XY\right]-E\left[X\right]E\left[Y\right]
\end{aligned}$$

Where I defined $\sum_{j=1}^{M}p_{ij}=p_i^X$ as the marginal distribution of X and $\sum_{i=1}^{N}p_{ij}=p_j^Y$ as the marginal distribution of Y.

Correlation Coefficient

$$\rho\left(X,Y\right)=\frac{Cov\left(X,Y\right)}{\sigma_X\sigma_Y}$$

- Unlike covariance, correlation coefficient is scale-free
- Correlation coefficient will always lie between -1 and 1

Useful Results

Proposition: $E\left[X+Y\right]=E\left[X\right]+E\left[Y\right]$

Proof:

$$\begin{aligned}
E[X+Y] &= \sum_{j=1}^{M}\sum_{i=1}^{N}(x_i+y_j)\,p_{ij} \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N}x_i p_{ij} + \sum_{j=1}^{M}\sum_{i=1}^{N}y_j p_{ij} \\
&= \sum_{i=1}^{N}x_i \sum_{j=1}^{M}p_{ij} + \sum_{j=1}^{M}y_j \sum_{i=1}^{N}p_{ij} \\
&= \sum_{i=1}^{N}x_i p_i^X + \sum_{j=1}^{M}y_j p_j^Y \\
&= E[X] + E[Y]
\end{aligned}$$

Proposition: $Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$

Proof:

$$\begin{aligned}
Var[X+Y] &= \sum_{j=1}^{M}\sum_{i=1}^{N}\left((x_i+y_j) - E[X+Y]\right)^2 p_{ij} \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N}\left((x_i+y_j) - E[X] - E[Y]\right)^2 p_{ij} \text{ by previous proposition} \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N}\left((x_i - E[X]) + (y_j - E[Y])\right)^2 p_{ij} \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N}\left((x_i - E[X])^2 + 2(x_i - E[X])(y_j - E[Y]) + (y_j - E[Y])^2\right) p_{ij} \\
&= \sum_{j=1}^{M}\sum_{i=1}^{N}(x_i - E[X])^2 p_{ij} + \sum_{j=1}^{M}\sum_{i=1}^{N}(y_j - E[Y])^2 p_{ij} \\
&\quad +2\sum_{j=1}^{M}\sum_{i=1}^{N}(x_i - E[X])(y_j - E[Y])\,p_{ij} \\
&= Var[X] + Var[Y] + 2Cov(X,Y)
\end{aligned}$$

Independence

- Independence roughly means that the probability of an outcome associated with Y will be unrelated with the outcome associated with X, and vice versa
- Knowledge of realization of X does not help updating the probability assessment of Y
- Example: Two consecutive coin tossings

Comment: The formal definition of independence says that if X and Y are independent, their joint distribution can be written as the product of their marginal distributions. Equivalently, $\forall i, j,\ p_{ij} = p_i^X \cdot p_j^Y$.

Proposition: If X and Y are independent $E\left[X \cdot Y\right] = E\left[X\right] \cdot E\left[Y\right]$

Proof: Let X and Y be independent. Then $\forall i, j,\ p_{ij} = p_i^X p_j^Y$.

$$\begin{aligned} E\left[X \cdot Y\right] &= \sum_{j=1}^{M}\sum_{i=1}^{N} x_i y_j p_{ij} \\ &= \sum_{j=1}^{M}\sum_{i=1}^{N} x_i y_j p_i^X p_j^Y \\ &= \left(\sum_{j=1}^{M} y_j p_j^Y\right)\left(\sum_{i=1}^{N} x_i p_i^X\right) \\ &= E\left[Y\right] E\left[X\right] \end{aligned}$$

Proposition: $Cov\left(X, Y\right) = 0$

Proof:

$$\begin{aligned} Cov\left(X, Y\right) &= E\left[XY\right] - E\left[X\right] E\left[Y\right] \\ &= E\left[X\right] E\left[Y\right] - E\left[X\right] E\left[Y\right] \text{ from previous proposition} \\ &= 0 \end{aligned}$$

Estimation

- Means, variances, and covariances can be measured with certainty only if we know the exact probability distribution
- We cannot know the true value of means or variance in general
- We use the sample information to obtain the best possible estimates
- *Estimator*: A rule which will give a sample estimate for each and every possible sample (Because the sample itself is random, an estimator is a random variable)
- *Estimate*: A number

Some decision science

- Let b denote an estimator of an arbitrary parameter β, say expected value. What kind of properties do we usually want from b?

- *Lack of Bias*: We usually want the expected value of b to be equal to the parameter of interest β. We say that b is unbiased if so. We define

$$\text{Bias} = E[b] - \beta$$

- *Efficiency*: While lack of bias seems to be a desirable property, that alone does not say anything about the dispersion of the estimator about the true parameter. We usually want small dispersion. We say that an unbiased estimator b is efficient if its variance is the smaller than the variance of any other unbiased estimators.

- *Minimum MSE*: Sometimes we have to trade off bias and variance of estimators. It seems useful to minimize the mean squared error

$$MSE = E\left[(b-\beta)^2\right]$$

Proposition: $E\left[(b-\beta)^2\right]$ =Bias$^2 + Var(b)$

Proof:

$$\begin{aligned} E\left[(b-\beta)^2\right] &= E\left[b^2 - 2b\beta + \beta^2\right] \\ &= E\left[b^2\right] - 2\beta E[b] + \beta^2 \\ &= \underbrace{E\left[b^2\right] - (E[b])^2}_{Var(b)} + \underbrace{(E[b])^2 - 2\beta E[b] + \beta^2}_{\text{Bias}^2} \\ &= \text{Bias}^2 + Var(b) \end{aligned}$$

- *Consistency* (Large sample property): We would like b to be very close to β as sample size increases. We want, as the sample size increases, the probability that b differ from β will be small. We say that $\operatorname{plim} b = \gamma$ if, as sample size approaches infinity, the probability that $|b-\gamma|$ will be less than any arbitrary small number will approach 1. We say that b is consistent if $\operatorname{plim} b = \beta$. Usually, econometricians tend to be more concerned with consistency than with lack of bias

Estimation of μ_X

- Expected value is usually estimated by sample mean

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

- Sample mean is unbiased for μ_X

$$E\left[\overline{X}\right] = E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N}E\left[\sum_{i=1}^{N} X_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[X_i] = \frac{1}{N}\sum_{i=1}^{N}\mu_X = \frac{1}{N}N\mu_X = \mu_X$$

Estimation of σ_X^2 and Cov

- Sample variance

  $$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$

  is unbiased for variance

Proposition: $s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$ is unbiased.

Proof:

$$\begin{aligned}
s_X^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2 \\
&= \left(\frac{1}{N-1}\right)\sum_{i=1}^{N}\left(X_i^2 - 2\bar{X}X_i + \bar{X}^2\right) \\
&= \left(\frac{1}{N-1}\right)\left[\sum_{i=1}^{N}X_i^2 - 2\bar{X}\sum_{i=1}^{N}X_i + \sum_{i=1}^{N}\bar{X}^2\right] \\
&= \left(\frac{1}{N-1}\right)\left[\sum_{i=1}^{N}\left(X_i^2\right) - 2N\bar{X}^2 + N\bar{X}^2\right] \\
&= \left(\frac{1}{N-1}\right)\left[\sum_{i=1}^{N}X_i^2 - N\bar{X}^2\right]
\end{aligned}$$

This gives us:

$$\begin{aligned}
E\left(s_X^2\right) &= E\left[\left(\frac{1}{N-1}\right)\left[\sum_{i=1}^{N}X_i^2 - n\bar{X}^2\right]\right] \\
&= \left(\frac{1}{N-1}\right)E\left[\sum_{i=1}^{N}X_i^2 - n\bar{X}^2\right] \\
&= \left(\frac{1}{N-1}\right)\left(\sum_{i=1}^{N}E\left[X_i^2\right] - n\bar{X}^2\right) \\
&= \left(\frac{1}{N-1}\right)\left(NE\left[X^2\right] - NE\left[\bar{X}^2\right]\right) \\
&= \left(\frac{N}{N-1}\right)\left(E\left[X^2\right] - E\left[\bar{X}^2\right]\right)
\end{aligned}$$

Recall that:

$$\begin{aligned}
Var\left(X\right) &= E\left[X^2\right] - \left(E\left[X\right]\right)^2 \\
E\left[X^2\right] &= Var\left(X\right) + \left(E\left[X\right]\right)^2
\end{aligned}$$

And (shown below):

$$\begin{aligned}
\frac{Var\left(X\right)}{N} &= E\left[\bar{X}^2\right]-\left(E\left[X\right]\right)^2 \text{ since } \bar{X} \text{ is unbiased}\\
E\left[\bar{X}^2\right] &= \frac{Var\left(X\right)}{N}+\left(E\left[X\right]\right)^2
\end{aligned}$$

This gives us:

$$\begin{aligned}
E\left(s_X^2\right) &= \left(\frac{N}{N-1}\right)\left(Var\left(X\right)+\left(E\left[X\right]\right)^2-\frac{Var\left(X\right)}{N}-\left(E\left(X\right)\right)^2\right)\\
&= \left(\frac{N}{N-1}\right)\left(\frac{NVar\left(X\right)-Var\left(X\right)}{N}\right)\\
&= Var\left(X\right)
\end{aligned}$$

- More intuitive estimator

  $$\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\overline{X}\right)^2$$

  is unbiased

Comment: Actually, this estimator IS biased.

- The right denominator $N-1$ is sometimes called the degree of freedom
- Likewise, sample covariance is an unbiased estimator of covariance

  $$\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)$$

  and sample correlation is defined by

  $$\frac{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\overline{X}\right)^2}\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2}}$$

Law of Large Numbers

- The variance of the sample mean

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

  is equal to

  $$\frac{1}{N^2}\sum_{i=1}^{N}Var\left(X_i\right)=\frac{1}{N^2}\sum_{i=1}^{N}\sigma^2=\frac{\sigma^2}{N}.$$

- Therefore, the variance of the sample mean is negligible if N is sufficiently large. Note that $Var\left(\overline{X}\right)\to 0$ as $N\to\infty$
- Roughly speaking, the sample mean almost becomes a nonstochastic number with very small dispersion around its mean μ
- When the sample size is large enough, the sample mean is virtually equal to μ

Normal Distribution

- Symmetric bell-shaped density
- Fully characterized by mean and variance
- If two or more random variables are normally distributed, any weighted sum of these variables will be normally distributed.
- In particular, if $X_1, \ldots, X_N$ constitute a random sample from $N(\mu, \sigma^2)$, i.e., they are independent and identically distributed, then

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

  and therefore

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{N}} \sim N(0, 1)$$

- If $X \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} \Pr[\mu - 1.96\sigma < X < \mu + 1.96\sigma] &= .95 \\ \Pr[\mu - 2.57\sigma < X < \mu + 2.57\sigma] &= .99 \end{aligned}$$

- If $X \sim N(0, 1)$, then we say that X has a standard normal distribution

Chi-Square distribution

- The sum of the squares of N independently distributed standard normal random variables is distributed as chi square distribution with N degrees of freedom
- We sometimes write it as $X \sim \chi^2(N)$

t distribution

- Assume that $X \sim N(0, 1)$ and $Z \sim \chi^2(N)$
- Further assume that they are independent
- Then,

$$\frac{X}{\sqrt{Z/N}} \sim t(N)$$

- It has a t-distribution with N degrees of freedom

F distribution

- Assume that $X \sim \chi^2(N_1)$ and $Z \sim \chi^2(N_2)$
- Further assume that they are independent
- Then,

$$\frac{X/N_1}{Z/N_2} \sim F(N_1, N_2)$$

Confidence Interval and Hypothesis Test for μ

- Suppose that $X_1, \ldots, X_N \sim N(\mu, \sigma^2)$
- Suppose that σ^2 is known
- 95% confidence interval for μ is given by

$$\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

  This is because

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{N}} \sim N(0,1)$$

  and hence

$$\Pr\left[\left|\frac{\overline{X} - \mu}{\sigma/\sqrt{N}}\right| < 1.96\right] = .95$$

  or

$$\Pr\left[-1.96\frac{\sigma}{\sqrt{n}} < \mu - \overline{X} < 1.96\frac{\sigma}{\sqrt{n}}\right] = .95$$

  or

$$\Pr\left[\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < +1.96\frac{\sigma}{\sqrt{n}}\right] = .95$$

- For example, if $N = 100, \overline{X} = 3$, and $\sigma = 10$, then it would be 3 ± 1.96
- Confidence intervals can be used to test hypotheses. Suppose that the null hypothesis is given by $H_0 : \mu = 0$. Because the 95% confidence interval is given by $(1.04, 4.96)$, it is unlikely that the null hypothesis is true. We may therefore reject the null hypothesis.
- Throughout the course, we implicitly assume that the alternative hypothesis is two sided
- Formally, we reject the null (under 5% significance level) $H_0 : \mu = \mu_0$ if

$$\left|\frac{\overline{X} - \mu_0}{\sigma/\sqrt{N}}\right| \geq 1.96$$

  It can be verified that this procedure is equivalent to the intuitive procedure based on 95% confidence intervals

- Roughly speaking, the hypothesis testing strategy is based on following intuition: We believe that, if $\mu = \mu_0$, an estimator of μ, i.e., $\overline{X}$ should be close to μ_0. If not, there is a significant evidence against the null. Therefore, it is sensible to reject the null when $\left|\overline{X} - \mu_0\right|$ is large. Now, the problem is that, there is a noise in the estimation of μ. Therefore, our intuition has to be modified a little bit. If we believe that there is a large noise, then it may be sensible to accept the null even when $\left|\overline{X} - \mu_0\right|$ is reasonably large. The ratio $\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{N}}\right|$ provides a sensible compromise between these two conflicts. It is statistically convenient becase we know $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{N}}$ has a standard normal distribution.

- When σ is unknown, which is a more realistic situation, we reject instead when

$$\left|\frac{\overline{X} - \mu_0}{s_X/\sqrt{N}}\right| \geq t_c\left(N-1\right)$$

  where $t_c\left(N-1\right)$ is the critical value from $\left(N-1\right)$. The ratio $\left|\frac{\overline{X}-\mu_0}{s_X/\sqrt{N}}\right|$ is called the t-statistic, and has a $t\left(N-1\right)$ distribution under the null

- As in normal distribution based confidence interval, t-distribution based inference is based on the fact that

$$\frac{\overline{X} - \mu}{s_X/\sqrt{N}} \sim t\left(N-1\right)$$

- Finding the critical value: See Table 3, designed for two-tail tests. It reports the probability that the t-value will exceed the number in the table in absolute value. Usually, we only care about the column under .05 (= usual significance level). The title "Percentiles..." is actually a misnomer (In some other books, we would want to look at the column under .025).