Unit II (Ch. 3-5)

ECON 3406

Dr. Josh Martin

Intro to probability

- The **probability** of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times
    - Probability is defined as a **proportion**, and it always takes values between 0 and 1 (or displayed as a percentage)

Law of large numbers

- Example: rolling a die many times
  - Let $\hat{p}_n$ be the proportion of outcomes that are 1 after the first n rolls
  - As the number of rolls increases, $\hat{p}_n$ will converge to the probability of rolling a 1, $p = \frac{1}{6}$
- The tendency of $\hat{p}_n$ to stabilize around p is described by the **Law of Large Numbers**

Law of large numbers

Law of large numbers

- LLN: As more observations are collected, the proportion $\hat{p}_n$ of occurrences with a particular outcome converges to the probability p of that outcome
- Occasionally the proportion will veer off from the probability and appear to "defy" the Law of Large Numbers
  - However, these deviations become smaller as the number of rolls increases

Disjoint (or mutually exclusive) outcomes

- Two outcomes are called disjoint (or mutually exclusive) if they cannot both happen
- For instance, if we roll a die, the outcomes 1 and 2 are disjoint since they cannot both occur
- On the other hand, the outcomes 1 and "rolling an odd number" are not disjoint since both occur if the outcome of the roll is a 1
- Calculating the probability of disjoint outcomes is easy – add their separate probabilities

$$P(1 \text{ or } 2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Addition Rule of Disjoint Outcomes

- If there are many disjoint outcomes $A_1, \ldots, A_k$, then the probability that one of these outcomes will occur is:

$$P(A_1) + P(A_2) + \cdots + P(A_k)$$

Disjoint outcomes

- Data scientists rarely work with individual outcomes and instead consider **sets** or **collections of outcomes**
- Let **A** represent the event where a die roll results in 1 or 2 and **B** represent the event that the die roll is a 4 or a 6
  - A = {1, 2}; B = {4, 6}
  - These sets are commonly called events
- Because A and B have no elements in common, they are disjoint events

$$P(\text{A or B}) = P(A) + P(B) = \frac{2}{6} + \frac{2}{6} = \frac{4}{6} = \frac{2}{3}$$

Reminder on card decks

- The 52 cards are split into four suits: club, diamond, heart, spade
  - Each suit has its 13 cards labeled: 2, 3, …, 10, jack, queen, king, and ace
  - Thus, each card is a unique combination of a suit and a label
  - The 12 cards represented by the jacks, queens, and kings are called face cards
  - The cards that are diamond and heart are typically colored red while the other two suits are typically colored black

Probabilities when events are not disjoint

- What is the probability that a randomly selected card is a diamond?
  - $P(\diamondsuit) = \frac{13}{52} = \frac{1}{4} = 0.25$
- What is the probability that a randomly selected card is a face card?
  - $P(\text{face}) = \frac{12}{52} \approx 0.23$
- What is the probability that a randomly selected card is a face card or a diamond?
  - $P(\text{face or } \diamondsuit) = P(\text{face}) + P(\diamondsuit) - P(\text{face and } \diamondsuit)$
  - $P(\text{face or } \diamondsuit) = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52} \approx 0.42$

General Addition Rule

- If A and B are any two events, disjoint or not, then the probability that at least one of them will occur is:

$$P(\text{A or B}) = P(A) + P(B) - P(\text{A and B})$$

- where $P(\text{A and B})$ is the probability that both events occur
- If A and B are disjoint, this implies $P(\text{A and B}) = 0$

Probability distribution

- A probability distribution is a representation of all disjoint outcomes and their associated probabilities
  - The outcomes listed must be disjoint
  - The probabilities must total 1

Probability distribution

Independence

- Two processes are **independent** if knowing the outcome of one provides no useful information about the outcome of the other
  - For instance, flipping a coin and rolling a die are two independent processes – knowing the coin was heads does not help determine the outcome of a die roll
  - On the other hand, stock prices usually move up or down together, so they are not independent

Multiplication rule for independent processes

- If A and B represent events from two different and independent processes, then the probability that both A and B occur can be calculated as the product of their separate probabilities:

$$P(\text{A and B}) = P(A) \times P(B)$$

- Similarly, if there are k events $A_1, \ldots, A_k$ from k independent processes, then the probability they all occur is:

$$P(A_1) \times P(A_2) \times \ldots \times P(Ak)$$

Multiplication rule for independent processes

- If we roll two dice, what is the probability that you roll two 1s?

$$P(1 \text{ and } 1) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

- 1/6th of the first rolls are a 1 and 1/6th of those times where the first roll is a 1 the second roll is also a 1

Multiplication rule for independent processes

```r
set.seed(1)
x <- data.frame(
  dice1 = sample(1:6, 10000, replace = TRUE),
  dice2 = sample(1:6, 10000, replace = TRUE)
)

x$ones <- ifelse(x$dice1 == 1 & x$dice2 == 1, 1, 0)
mean(x$ones)
```

```
[1] 0.0283
```

```r
1/36
```

```
[1] 0.02777778
```

```r
nrow(x)
```

```
[1] 10000
```

Example: Photo Classification

- Data scientists have been working to improve a classifier for whether the photo is about fashion or not
  - The “photo classify” data set represents a sample of 1822 photos from a photo sharing website
- Each photo gets two classifications:
  - “mach_learn” and gives a classification from a machine learning (ML) system of either “pred_fashion” or “pred_not”
  - Each of these 1822 photos have also been classified carefully by a team of people, which we take to be the source of truth which takes values “fashion” and “not”

Example: Photo Classification

	Fashion	Not	Total
$\hat{}$ Fashion	197	22	219
$\hat{}$ Not	112	1491	1603
Total	309	1513	1822

Cheat sheet

Machine_Learning	Fashion	Not	Total
^ Fashion	197 (90%) [63.8%]	22 (10%) [1.5%]	219 (100%) [12%]
^ Not	112 (7%) [36.2%]	1491 (93%) [98.5%]	1603 (100%) [88%]
Total	309 (17%) [100%]	1513 (83%) [100%]	1822 (100%) [100%]

Marginal Probabilities

	Fashion	Not	Total
$\hat{}$ Fashion	197	22	219
$\hat{}$ Not	112	1491	1603
Total	309	1513	1822

- What is the probability that the machine algorithm will classify a picture as being about fashion?
  - $\frac{219}{1822} \approx 12\%$
- **Marginal probabilities** are the probabilities based on a single variable without regard to any other variables

Joint Probabilities

	Fashion	Not	Total
$\widehat{\text{Fashion}}$	197	22	219
$\widehat{\text{Not}}$	112	1491	1603
Total	309	1513	1822

- What is the probability that the machine algorithm will classify a picture as being about fashion *and*…
  - what is the probability that picture is about fashion?
  - $P(\widehat{Fashion} \text{ and } Fashion) = \frac{197}{1822} \approx 10.8\%$

Joint Probabilities

```
             joint_outcome prob
1 pred_fashion and fashion 10.8
2     pred_not and fashion  6.1
3     pred_fashion and not  1.2
4         pred_not and not 81.8
```

- What is the probability that the machine algorithm will classify a picture as being about fashion *and*…
  - what is the probability that picture is about fashion?
  - $P(\hat{Fashion} \text{ and } Fashion) = \frac{197}{1822} \approx 10.8\%$
- A probability of outcomes for two or more variables or processes is called a **joint probability**
  - note: the above table is a probability distribution

Conditional Probability

- **Conditional probability** is a measure of the probability of an event occurring, given that another event (by assumption, presumption, assertion or evidence) is already known to have occurred
- There are two parts to a conditional probability
  - the outcome of interest (A)
  - and the condition (B)
- It is useful to think of the condition as information we know to be true
- Notation: **P(A|B)** = “the probability that A is true given that B is true”

Conditional Probability

	Fashion	Not	Total
$\widehat{\text{Fashion}}$	197	22	219
$\widehat{\text{Not}}$	112	1491	1603
Total	309	1513	1822

- Conditional upon the machine learning prediction that the picture is about fashion, what is the probability that the picture is truly about fashion?
  - $P(Fashion \mid \widehat{Fashion}) = \frac{197}{219} \approx 90\%$

Conditional Probability

	Fashion	Not	Total
$\widehat{\text{Fashion}}$	197	22	219
$\widehat{\text{Not}}$	112	1491	1603
Total	309	1513	1822

- Takeaways: this measures **model precision**
  - When the algorithm predicts that a picture is about fashion, it is correct about 90% of the time
  - In practice, this means that the model's positive predictions are usually trustworthy – false positives are relatively rare

Conditional Probability

	Fashion	Not	Total
$\widehat{\text{Fashion}}$	197	22	219
$\widehat{\text{Not}}$	112	1491	1603
Total	309	1513	1822

- Conditional upon the picture is truly being about fashion, what is the probability that the machine learning estimation predicted that the picture is about fashion?
  - $P(\widehat{Fashion} \mid Fashion) = \frac{197}{309} \approx 63.8\%$

Conditional Probability

	Fashion	Not	Total
$\hat{\text{Fashion}}$	197	22	219
$\hat{\text{Not}}$	112	1491	1603
Total	309	1513	1822

- Takeaways: this measures **model sensitivity**
  - Out of all the pictures that are truly about fashion, the model successfully identifies about 64% of them
  - In practice, this means the model misses a fair number of true fashion images – false negatives are fairly common

Conditional Probability

- The model is conservative: when it does say “fashion,” it’s usually right (high precision)
- But it is also incomplete: it fails to flag a sizable share of true fashion images (moderate sensitivity)
- So the algorithm is good at being correct when it predicts fashion, but not great at catching all fashion-related content

Conditional Probability

- What if we didn't have count data, but instead only had probabilities?

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

- $P(Fashion \mid \hat{Fashion})$
    - $P(\hat{Fashion} \mid \hat{Fashion}) = \frac{P(\hat{Fashion} \text{ and } \hat{Fashion})}{P(\hat{Fashion})} = \frac{0.108}{0.120} = 90\%$

Conditional Probability: Example II

	Vaccinated	Not	Total
Lived	238	5136	5374
Died	6	844	850
Total	244	5980	6224

- The smallpox data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston
- Doctors hypothesized that inoculation, which involves exposing a person to the disease in a controlled form, could reduce the likelihood of death

Conditional Probability: Example II

	Vaccinated	Not	Total
Lived	3.8%	82.5%	86.3%
Died	0.1%	13.6%	13.7%
Total	3.9%	96.1%	100%

- What is the the probability that a randomly selected person who was (not) inoculated died from smallpox?
  - $P(Died \mid Vaccinated) = \frac{0.001}{0.039} \approx 2.6\%$
  - $P(Died \mid Not) = \frac{0.136}{0.961} \approx 14.1\%$

General Multiplication Rule

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(A \text{ and } B) = P(A \mid B) \times P(B)$$

Conditional Probability: Example II

	Vaccinated	Not	Total
Lived	238	5136	5374
Died	6	844	850
Total	244	5980	6224

- How could we compute the probability that a resident was not inoculated and lived?
  - $P(\text{Lived and Not}) = P(Lived \,|\, Not) \times P(Not)$
  - $P(\text{Lived and Not}) = P(Lived \,|\, Not) \times P(Not) = \frac{5136}{5980} \times \frac{5980}{6224} \approx 82.5\%$

Conditional Probability: Example II

	Vaccinated	Not	Total
Lived	238	5136	5374
Died	6	844	850
Total	244	5980	6224

- How could we compute the probability that a resident was not inoculated and lived?
  - $P(\text{Lived and Not}) = P(Not \mid Lived) \times P(Lived)$
  - $P(\text{Lived and Not}) = P(Not \mid Lived) \times P(Lived) = \frac{5136}{5374} \times \frac{5374}{6224} \approx 82.5\%$

Conditional Probability: Example II

Point Estimates

- Suppose a poll suggests the U.S. President's approval rating is 45%.
  - We would consider 45% to be a **point estimate** of the approval rating we might see if we surveyed the entire population.
  - This entire-population proportion is referred to as the **parameter** of interest.
  - When the parameter is a proportion, it is often denoted by p, while the sample proportion is denoted $\hat{p}$.
  - Unless we survey every individual in the population, p remains unknown, and we use $\hat{p}$ as our estimate of p.

Statistical Error

- The difference between a poll's result and the true parameter is called the **error** in the estimate.
  - In general, error has two components: **sampling error** and **bias**.
- **Sampling error** (or sampling uncertainty) describes how much an estimate will vary from one sample to the next.
  - The size of the sample, n, is critical for quantifying this error.
- **Bias** refers to a systematic tendency to over- or under-estimate the true population value.
  - Example: a student poll on support for a new college stadium would likely overstate support among the full population.

Standard Errors

What They Measure

- **Standard Deviation (SD):**
  - Measures variability of individual data points around the mean.
  - Describes how spread out the data are within the sample or population.
  - Example: “On average, test scores vary ± 10 points around the mean.”

Standard Errors

What They Measure

- **Standard Error (SE):**
  - Measures variability of a sample statistic (like the mean) across repeated samples.
  - $SE_{\hat{p}} = \frac{\sigma}{\sqrt{n}}$ and $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
  - Example: “If we repeatedly sampled students, the sample mean would vary by about ± 2 points.”

Standard Errors

Use Cases

- **Use SD when:**
  - Describing the spread of raw data in one sample.
  - Comparing individual variability.
  - Reporting descriptive statistics (e.g., average income with spread).
- **Use SE when:**
  - Estimating the precision of a sample statistic.
  - Showing uncertainty around an estimate.
  - Constructing confidence intervals or conducting hypothesis tests.

Standard Errors

Interpretation Tradeoffs

- **Standard Deviation (SD):**
  - Reflects the actual variability in the dataset.
  - Does not shrink with larger sample size.
  - Intuitive for describing: “How different are individuals from each other?”

Standard Errors

Interpretation Tradeoffs

- **Standard Error (SE):**
  - Shrinks as sample size increases (because of $\sqrt{n}$ in the denominator).
  - More relevant for inference: “How precisely do we know the mean?”
  - Can be misleading if reported instead of SD: large datasets produce small SEs even when the data are highly variable.

Standard Errors

Reporting in Practice

- To describe your sample → report **mean ± SD**.
- To estimate for the population → report **mean ± SE**.
- **Rule of Thumb:**
  - Use **SD** to describe data.
  - Use **SE** to describe the precision of an estimate.

Confidence Intervals

- The sample proportion $\hat{p}$ provides one plausible value for the population proportion p.
  - However, $\hat{p}$ is not perfect and has some standard error associated with it.
- When reporting an estimate of p, it is better practice to provide a **range of plausible values** rather than just a single point estimate.
  - Using only a point estimate is like fishing in a murky lake with a spear: we may throw the spear where we think a fish is, but we'll often miss.
  - Using an interval is like casting a net in that area: we have a much better chance of catching the fish.

Confidence Intervals

- A **confidence interval** is like fishing with a net—it represents a range of plausible values where we are likely to find the population parameter.
- Our sample proportion $\hat{p}$ is the most plausible value of p, so it makes sense to build a confidence interval around this point estimate.
  - The **standard error** guides how wide the confidence interval should be.

Margin of Error

- The standard error is the standard deviation of the point estimate.
  - In a normal distribution, about 95% of the data lie within 1.96 standard deviations of the mean.
  - By the same logic, we can construct a 95% confidence interval by extending 1.96 standard errors on either side of the point estimate:

$$\text{point estimate} \pm 1.96 \times SE$$

- Margin of error = $z \times SE$

Variability of the Point Estimate

- Suppose the true proportion of American adults who support the expansion of solar energy is $p = 0.88$.
  - If we surveyed 1,000 American adults, our estimate would not equal 0.88 exactly. How close would we expect the sample proportion $\hat{p}$ to be?
- To think about this:
  - Imagine 250 million slips of paper (the number of American adults in 2018), with "support" written on 88% and "not" on 12%.
  - Draw 1,000 slips at random to represent our sample.
  - Compute the fraction of "support" responses in the sample.

Simulation Analysis

```r
# Set population size and true proportion of support
pop <- 250000       # total population size
p <- 0.88           # true population proportion that
loops <- 1000       # number of simulations (how many

# Create an empty list to store results
z <- list()
n <- 250            # sample size for each poll

# Run simulation "loops" times
for(i in 1:loops){

  # Create population: 1 = support, 0 = not support
  support <- rep(1, pop * p)           # 88% supporte
  not <- rep(0, pop * (1 - p))         # 12% non-supp

  # Randomize order (so we can sample randomly)
  set.seed(i)                          # seed for rep
  s <- data.frame(
```

Confidence Intervals

- But what does “95% confident” really mean?
  - Imagine we repeatedly took many samples and built a 95% confidence interval from each one.
  - About 95% of those intervals would contain the true parameter p.

```
mean(ifelse(df$point_est < lower_bound_se, 1,
       ifelse(df$point_est > upper_bound_se, 1, 0)))
```

```
[1] 0.06
```

Sample Size and Variance

n = 25

n = 250

n = 2500

SE vs. SD

n = 25
n = 250
n = 2500

Why Standard Errors?

- In the real world, we don't observe the DGP or all possible events!
  - Thus, the standard deviation is not knowable

```
se <- mean(df$sd) / sqrt(n)
se
```

```
[1] 0.02060871
```

```
sqrt(p*(1-p)/n)
```

```
[1] 0.02055237
```

Why 95% Confidence Intervals?

- **95% is the most common standard** in applied work
  - Balances **precision** (narrower intervals) and **certainty** (coverage probability)
- A 95% CI means:
  - If we repeated sampling many times, ~95% of those intervals would contain the true parameter
- Wider confidence intervals (like 99%) reduce Type I errors, but at the cost of greater imprecision

95% vs. 99%: Error Tradeoffs

- **Type I Error (false positive):**
  - More likely with a 95% interval than a 99% interval
  - Example: claiming a treatment works when it doesn't
- **Type II Error (false negative):**
  - More likely with a 99% interval than a 95% interval
  - Example: missing a real effect because the interval is too wide
- **Key tradeoff:**
  - 95% intervals → more power, but slightly higher risk of false positives
  - 99% intervals → more caution, but higher chance of missing true effects

Hypothesis Testing: The Basics

- **Null hypothesis (H_0):** a default claim (e.g., “no effect” or “no difference”)
- **Alternative hypothesis (H_A):** the competing claim (e.g., “there is an effect” or “a difference exists”)
- Goal: use data to assess whether we have *enough evidence* to reject H_0

Analogy: Innocent Until Proven Guilty

- Think of H_0 as **"the defendant is innocent"**
- H_A as **"the defendant is guilty"**
- We require **strong evidence** before rejecting H_0
- Just as in court:
  - Failing to reject H_0 ≠ proving innocence
  - It means evidence was insufficient to declare guilt

Hypothesis Testing and Double Negatives

- Grammatically: “Lack of evidence is **not** evidence of absence”
- In stats:
  - Failing to reject H_0 does not prove H_0 is true
  - It only means we didn’t find strong enough evidence against it
- Key mindset: hypothesis tests never “prove,” they only **reject** or **fail to reject**

Statistical Power: What It Is

- **Power analysis** is used to determine the *minimum sample size* needed for an experiment
  - Helps to ensures results are reliable and not simply due to chance
- Statistical power, more broadly, refers to the probability of detecting an effect (i.e. rejecting the null hypothesis) given that the effect exists

Statistical Power: β

- If H_A is true, there are only two possible outcomes:
  - We correctly reject H_0 (a success).
  - We fail to reject H_0 (a miss, i.e. Type II error).
- The probabilities of these two outcomes must add to 1.
  - Probability of miss = β.
  - Probability of success = $1 - \beta$.

$$Power = P(\text{Reject } H_0 \mid H_A \text{true}) = 1 - \beta$$

Statistical Power

- **Significance level (α):**
  - Probability of rejecting H_0 when H_0 is actually true (Type I Error)
- **Power ($1 - \beta$):**
  - Probability of rejecting H_0 when H_1 is true
  - Here, β is the probability of a Type II Error

	Reject H_0	**Fail to Reject H_0**
H_0 is True	α (Type I)	$1 - \alpha$
H_1 is True	$1 - \beta$ (Power)	β (Type II)

Trade-offs in Power

- Stricter significance levels → smaller rejection regions
  - Reduce **Type I errors**
  - Increase risk of **Type II errors** (lower power)
- Designing studies requires balancing:
  - Caution against false positives
  - Adequate sensitivity to detect real effects

Why Power Matters

- An **underpowered study**:
  - Likely inconclusive, unable to distinguish between H_0 and H_A
- An **overpowered study**:
  - Detects even trivial differences
  - Risks wasting time, money, and resources
- Both extremes can undermine the value of research

Broader Implications

- Too many underpowered studies:
  - Increase in false positives among published results
  - Contributes to the **replication crisis** in science
- Excessive demands for power:
  - Resource waste
  - Ethical issues (e.g., unnecessary use of animal subjects)

What Determines Power?

Power depends on three main factors:

1. **The test itself** and the significance level (α)
2. **The magnitude of the effect** being studied
3. **Sample size and variability** in the data

- Researchers can control these factors when designing experiments

Power Analysis: The Math

- Recall that: **Margin of error =** $z \times SE$
- Thus, if we want our margin of error to be ≈ 0.01 at a 95% significance level:

$$0.01 = 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

$$0.01^2 = 1.96^2 \times \frac{p(1-p)}{n}$$

$$n = \frac{1.96^2 \times p(1-p)}{0.01^2}$$

Power Analysis: The Math

$$n = \frac{1.96^2 \times p(1-p)}{0.01^2}$$

```
1 (1.96^2 * p * (1 - p)) / 0.01^2
```

```
[1] 4056.73
```

Simulation Analysis: Pt. II

```r
# Set population size and true proportion of support
pop <- 250000       # total population size
p <- 0.88           # true population proportion that
loops <- 1000       # number of simulations (how many

# Create an empty list to store results
z <- list()
n <- 4057           # sample size for each poll

# Run simulation "loops" times
for(i in 1:loops){

  # Create population: 1 = support, 0 = not support
  support <- rep(1, pop * p)           # 88% supporte
  not <- rep(0, pop * (1 - p))         # 12% non-supp

  # Randomize order (so we can sample randomly)
  set.seed(i)                          # seed for rep
  s <- data.frame(
```

Simulation Analysis: Pt. II

The results

```
1 se * 1.96
```

```
[1] 0.01000569
```

```
1 sd(df$point_est) * 1.96
```

```
[1] 0.01008457
```