Probability Concepts Review

2017.3.6 - 2017.3.8

Probability

- Lowest value: 0
  - Chance of event that is impossible
- Highest value: 1 (or 100%)
  - Chance of event that is certain
- If an event has chance 70%, then the chance that it doesn't happen is
  - 100% - 70% = 30%
  - 1 - 0.7 = 0.3

Equally Likely Outcomes

Assuming all outcomes are equally likely, the chance of an event A is:

$$P(A) = \frac{\text{number of outcomes that make A happen}}{\text{total number of outcomes}}$$

Multiplication Rule

Chance that two events *A* and *B* both happen

= P(*A* happens) x P(*B* happens given that *A* has happened)

- The answer is *less than or equal to* each of the two chances being multiplied
- The more conditions you have to satisfy, the less likely you are to satisfy them all

Addition Rule

If event *A* can happen in *exactly one* of two ways, then

$$P(A) = P(\text{first way}) + P(\text{second way})$$

- The answer is *greater than or equal to* the chance of each individual way

Sampling

- Deterministic sample:
  - Sampling scheme doesn't involve chance
- Probability sample:
  - Before the sample is drawn, you have to know the selection probability of every group of people in the population
  - Not all individuals have to have equal chance of being selected

Estimation

Making conclusions based on data in random samples

Bias

- **Biased estimate:** On average across all possible samples, the estimate is either too high or too low.
- Bias creates a systematic error in one direction.
- Good estimates typically have low bias.

Variability

- The degree to which the value of an estimate **varies** from one sample to another.
- High variability makes it hard to estimate accurately.
- Good estimates typically have low variability.

Distribution of a Statistic

Statistic: A quantity computed for a particular sample

Distribution: The chance of each outcome of sampling

Sampling distribution: Chance of each value of a statistic (computed from all possible samples)

Also known as the *probability distribution of the statistic*

Empirical distribution: Observations of a statistic (computed from some samples drawn at random)

Law of Averages

If a chance experiment is repeated many times, independently and under the same conditions, then the proportion of times that an event occurs gets closer to the theoretical probability of the event

Large Random Samples

If the sample size is large, then the empirical distribution of a uniform random sample resembles the distribution of the population, with high probability

Computing Distance

Total Variation Distance (TVD):

- For each category, compute the difference in proportions between two distributions
- Take the absolute value of each difference
- Sum & divide by 2

Chi Squared (χ^2 *Optional*):

- For each category, compute the difference in proportions between two distributions
- Square each difference and divide by the first proportion
- Sum & multiply by sample size

Testing a Hypothesis

Step 1: The Hypotheses

- A test chooses between two views of how data were generated
- *Null hypothesis* proposes that data were generated at random
- *Alternative hypothesis* proposes some effect other than chance

Step 2: The Test Statistic

- A value that can be computed for the data and for samples

Step 3: The Sampling Distribution of the Test Statistic

- What the test statistic might be if the null hypothesis were true
- Approximate the sampling distribution by an empirical distribution

Conclusion of a Test

Resolve choice between null and alternative hypotheses

- Compare observed test statistic to its empirical distribution under the null hypothesis
- If the observed value is **consistent** with the distribution, then the test *does not* support the alternative hypothesis

Whether a value is consistent with a distribution:

- A visualization may be sufficient
- Convention: The observed significance level (P-value)

Observed Significance Level

P-Value: The chance, under the null hypothesis, that the test statistic is equal to the value that was observed or is even further in the direction of the alternative.

Statistically Significant: The P-value is less than 5%

Highly Statistically Significant: The P-value is less than 1%

Final words...

Don’t stress.

Do your best.

Forget the rest.

You can do it!