Probability and Statistics Refresher

Kevin Johnson

August 29, 2014

1 Probability Introduction

1.1 Definitions

First I'll give some definitions.

- **Population**: a collection of objects (e.g. Atlanta population, Georgia Tech students).
- **Sample**: subset of the population.
- **Random Sample**: randomly selected sample.
- **Sample Size**: number of objects in the sample.
- **Experiment**: measure characteristics of the sample/population.
- **Outcome**: possible values of the measurements.
- **Sample Space**: space of all possible outcomes (generally denoted by Ω).

For example, let's say our experiment is to roll two dice simultaneously. Our sample space is $\Omega = \{(1,1),(1,2),\ldots,(1,6),\ldots,(5,6),(6,6)\}$. One possible outcome is $(6,6)$, and there are $6 \cdot 6 = 36$ total possible outcomes.

An **Event** is any collection of sample outcomes. This can be a simple collection of outcomes or a collection described as unions and intersections of other events. A **Probability** is the likelihood of the occurrence of an outcome in the sample space.

1.2 Fundamental Properties

1. $0 \leq P(A) \leq 1$ for each event A.
2. $P(\Omega) = 1$

2 Unions and Intersections

Unions and intersections are used to denote common words like "AND" and "OR". Formally:

$$P(A \cup B) = P(\text{Outcome contained in Event A OR Event B}) \quad (1)$$
$$= P(A) + P(B) - P(A \cap B) \quad (2)$$
$$P(A \cap B) = P(\text{Outcome contained in Event A AND Event B}) \quad (3)$$
$$= P(A) \cdot P(B) \text{ (if A and B are independent)} \quad (4)$$

3 Counting

Often in probability you must count the total number of possible events in order to calculate the probability of an event occurring. There is an entire field dedicated to this called combinatorial analysis, but I will provide some basic rules.

3.1 Fundamental Principle of Counting (Multiplication Principle)

If the number of outcomes of experiment 1 is m, and the number of outcomes of experiment 2 is n, then the total number of outcomes for the two experiments is $m \cdot n$.

3.2 Permutations (Order is Important)

A permutation of objects occurs when objects are arranged so that order is important. The formal definition of a permutation (an arrangement of r out of n distinct objects where order is important) is:

$$P_{n,r} = \frac{n!}{(n-r)!} \quad (5)$$

3.3 Combinations (Order is Unimportant)

A combination of objects occurs when objects are selected and the order of arrangements is not important. The formal definition of a combination (selecting r out of n distinct objects where the order is unimportant) is:

$$C_{n,r} = \frac{n!}{r!(n-r)!} \tag{6}$$

4 Conditional Probability

A conditional probability is the probability that event A happens given that event B is known to happen.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{7}$$

Let's say we draw a card at random from a standard deck of 52 cards. We can use the above formula to calculate the probability of the following scenarios:

- Probability that the card is the Ace of Hearts (A) given that the card is Red (B).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{\frac{1}{52}}{\frac{1}{2}} = \frac{1}{26} \tag{8}$$

- Probability that the card is a King (A) given that the card is Red (B).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{2}{52}}{\frac{26}{52}} = \frac{1}{13} \tag{9}$$

In the last problem, note that the probability of getting a King is equal to the probability of getting a King given the card is Red ($P(\text{King}) = 4/52 = 1/13$). The information about the card color did not effect the probability of drawing a King. Therefore, the two events are **Independent**. Formally, events A and B are independent if:

$$P(A|B) = P(A) \tag{10}$$

$$P(A \cap B) = P(A)P(B) \tag{11}$$

This leads to one of the most important equations in probability, and the basis for all of Bayesian Statistics, **Bayes' Theorem**.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{12}$$

5 Random Variables

A **Random Variable** is the number assigned to the (random) outcome of an experiment in order to summarize its meaning. For example, let's say we have a deck of 52 cards and we randomly draw 5 cards. We can define a random variable X as the number of aces in the hand. X can take on a value of 0, 1, 2, 3, or 4. We can determine the probability that X will equal one of these values.

$$P(X = 0) = \left(\frac{48}{52}\right)\left(\frac{47}{51}\right)\left(\frac{46}{50}\right)\left(\frac{45}{49}\right)\left(\frac{44}{48}\right) \tag{13}$$

There are two types of random variables: **Discrete** and **Continuous**.

5.1 Discrete Random Variables

Discrete random variables have a finite or countably infinite range of outcomes, and we can write the outcomes as $\{X_1, X_2, \dots\}$. The set of probabilities associated to outcomes in the sample space is known as the **Probability Mass Function (PMF)**.

$$P_X(x_i) = P(X = x_i) \text{ for all possible values } x_i \tag{14}$$

The PMF of a random variable X defined as the result of one roll of a die is:

$$P_X(x) = P(X = x) = \begin{cases} \frac{1}{6} & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{o.w.} \end{cases} \tag{15}$$

The **Cumulative Distribution Function (CDF)** of a random variable is defined as:

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) \tag{16}$$

Note that $0 \leq F(x) \leq 1$ and $F(x)$ is always increasing as x increases. The CDF of the die example from before is:

$$P_X(x) = P(X = x) = \begin{cases} \frac{x}{6} & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{o.w.} \end{cases} \tag{17}$$

5.2 Continuous Random Variables

A continuous random variable X can be any real value in a given interval, so outcomes are not "countable". They are defined by a **Probability Density Function (PDF)** of X (known as $f(x)$) where $f(x) \geq 0$ does not represent a probability. The PDF of X at any point x is equal to 0. Rather than defining probabilities of individual points, PDFs define probabilities for intervals.

$$P(a \leq X \leq b) = \int_a^b f(t)dt \tag{18}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{19}$$

The CDF can be expressed as a function of the PDF:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \tag{20}$$

For example, let $F(x) = x^2/16$ from $0 \leq x \leq 4$. Then:

$$P(X \leq 2) = F(2) = 4/16 = 1/4 \tag{21}$$

$$P(1 \leq X \leq 3) = P(X \leq 3) - P(X \leq 1) \tag{22}$$

$$= F(3) - F(1) \tag{23}$$

$$= 9/16 - 1/16 = 1/2 \tag{24}$$

$$f(x) = \frac{d}{dx} F_X(x) = \frac{2x}{16} = \frac{x}{8} \tag{25}$$

6 Expectations

The **Expectation** of a random variable X represents the average value we expect to see from X. It is an average of all possible values of X weighted by

the probability of each of those values occurring. Formally,

$$E[X] = \begin{cases} \sum_x x P_X(x) & \text{if } X \text{ is discrete.} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{if } X \text{ is continuous.} \end{cases} \tag{26}$$

For example, let X be the outcome of a roll of a single die. The PMF of X has already been defined in Equation 15.

$$E[X] = \sum_x x P_X(x) = \sum_{i=1}^{6} \frac{i}{6} \tag{27}$$

$$= \frac{1}{6}(1+2+3+4+5+6) = 3.5 \tag{28}$$

Let X be a random variable with PDF $f_X(x) = \frac{1}{5}$ for $x \in [0, 5]$.

$$E[X] = \int_0^5 x \frac{1}{5} dx = 2.5 \tag{29}$$

7 Variability

The **Variance** of a random variable X is defined as the second central moment of X.

$$Var(X) \equiv \sigma_X^2 = E\left[(X - E[X])^2\right] = E[X^2] - E[X]^2 \tag{30}$$

Let X be the outcome of a roll of a single die.

$$E[X] = 3.5 \tag{31}$$

$$E[X^2] = \sum_{i=1}^{6} x_i^2 P_X(x) = \sum_{i=1}^{6} \frac{i^2}{6} \tag{32}$$

$$= \frac{1}{6}(1+4+9+16+25+36) = 15.167 \tag{33}$$

$$\sigma_X^2 = 15.167 - (3.5)^2 = 2.9167 \tag{34}$$

Let X be a random variable with PDF $f_X(x) = \frac{1}{5}$ for $x \in [0, 5]$.

$$E[X] = 2.5 \tag{35}$$

$$E[X^2] = \int_0^5 x^2 \frac{1}{5} dx = 8.33 \tag{36}$$

$$\sigma_X^2 = 8.33 - (2.5)^2 = 2.08 \tag{37}$$

8 Functions of Random Variables

It is often helpful to be able to know what happens when two random variables are added to each other or multiplied by a constant. The following is a brief list of rules for common situations.

$$E[aX + b] = aE[X] + b \tag{38}$$
$$E[X + Y] = E[X] + E[Y] \tag{39}$$
$$Var(aX + b) = E[[(aX + b) - (aE[X] + b)]^2] \tag{40}$$
$$= E[a^2[X - E[X]]^2] = a^2 Var(X) \tag{41}$$
$$Var(X + Y) = E[(X + Y)^2] - E[X + Y]^2 \tag{42}$$
$$= Var(X) + Var(Y) + 2Cov(X, Y) \tag{43}$$

9 Common Discrete Probability Distributions

9.1 Binomial Distribution

Let's say you have n independent trials or events, two outcomes for every trial (e.g. success and failure), and a probability p of success that is the same for every trial. We can define a random variable X as the number of successes out of n trials. We say that X follows a Binomial distribution ($X \sim \text{Bin}(n, p)$).

To define the PMF, we know that $P(X = k) = P(k \text{ successes and } n - k \text{ failures})$. From counting rules, we know there are $\binom{n}{k}$ ways of choosing k successes from a group of k successes and $n - k$ failures. Therefore:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \text{ for } k \in \{0, 1, 2, \ldots, n\} \tag{44}$$

The Binomial distribution has a mean np and variance $np(1 - p)$. Common applications include the number of heads in n coin tosses and the number of defectives in a product shipment of size n.

For example, we have a true/false test with 20 questions. Let X be the number of correct answers out of 20. To get an A, you need to get at least 19 questions correct. What is the probability that somebody who guesses every

answer will get an A ($X \sim \text{Bin}(20, 0.5)$)?

$$P(\text{get an A}) = P(X \geq 19) = P(X = 19) + P(X = 20) \tag{45}$$

$$= \binom{20}{19}(0.5)^{19}(1-0.5)^1 + \binom{20}{20}(0.5)^{20}(1-0.5)^0 \tag{46}$$

$$= 0.00002 \tag{47}$$

9.2 Poisson Distribution

The Poisson distribution is mainly useful for counting randomly occurring events. It is defined as Poisson(λ) where λ is the rate of occurrence.

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ for } x \in \{0, 1, 2, \dots\} \tag{48}$$

The Poisson distribution has a mean λ and variance λ, a unique property among distributions.

For example, the number of cracks in a ceramic tile has a Poisson distribution with rate $\lambda = 2.4$. What is the probability that the tile has no cracks? What is the probability that the tile has 4 or more cracks?

$$P(X = 0) = e^{-\lambda} = 0.09 \tag{49}$$

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - F_X(3) \tag{50}$$

$$= 1 - (P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)) \tag{51}$$

$$= 0.22 \tag{52}$$

It is interesting to note that the Poisson distribution can be an approximation for the the Binomial distribution when n is large and p is small. The Binomial distribution is cumbersome to work with for large values of n, so we can approximate it in the following way:

$$\text{Bin}(n, p) \approx \text{Poisson}(np) \tag{53}$$

10 Common Continuous Distributions

10.1 Uniform Distribution

The Uniform distribution is a symmetric distribution where all intervals of the same length within the range of the random variable are equally probable. It is defined as $\mathrm{U}[a, b]$ where a and b are its minimum and maximum values, respectively.

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{o.w.} \end{cases} \tag{54}$$

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases} \tag{55}$$

The Uniform distribution has a mean $\frac{1}{2}(a+b)$ and variance $\frac{1}{12}(b-a)^2$.

The main application of the Uniform distribution is in random number generation. It is relatively easy for computers to generate random numbers from a Uniform distribution, and you can express most many other distributions as a simple combination of Uniform random numbers (often using the standard Uniform $\mathrm{U}[0, 1]$). This topic is explored in depth in Simulation.

10.2 Exponential Distribution

The Exponential distribution describes the time between events in a Poisson process (i.e. a process in which events occur continuously and independently at a constant average rate). It is defined as $\mathrm{Exp}(\lambda)$ where λ is, again, the rate of occurrence.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{o.w.} \end{cases} \tag{56}$$

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & \text{o.w.} \end{cases} \tag{57}$$

The Exponential distribution has a mean λ^{-1} and variance λ^{-2}.

The Exponential distribution has many unique and important properties, but none are more important than the memoryless property. Suppose that

$X \sim \text{Exp}(\lambda)$ is the lifetime of a car engine with $\lambda = 100000$ miles. If the engine has lasted 200000 miles, then intuitively we would assume it will soon fail. However, the memoryless property means that the probability that the engine lasts another 100000 miles is the same as the probability that the original engine will last 100000 miles. It does not matter how long you have been waiting for a given event to occur, if it follows an Exponential distribution then the time it until the next event is always $\text{Exp}(\lambda)$. Formally,

$$P(X > t + s | X > t) = P(X > s) \tag{58}$$

10.3 Normal Distribution

The only probability distribution more important than the Exponential distribution is the Normal distribution. Entire classes have been devoted to this distribution, but I will try to give you some basic facts, useful properties, and applications. It is defined as $(N)(\mu, \sigma^2)$.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for } -\infty < x < \infty \tag{59}$$

$$F(x) = \Phi(x) \tag{60}$$

There is no closed form expression for the CDF of a Normal distribution, instead we use the function $\Phi(x)$. There are tables that provide values for $\Phi(x)$, and every statistical software package has functions that will provide values for any x. Typically, Z is used to denote the standard Normal with mean 0 and variance 1. $\Phi(x)$ is a very important function in Statistics, so it's useful to lay out a few of it's properties:

- $\Phi_Z(0) = 1/2$
- $\Phi_Z(\infty) = 1$
- $\Phi_Z(z) = 1 - \Phi_Z(-z)$
- $P(Z \leq z) = 1 - P(Z \leq -z)$
- $P(Z \geq z) = P(Z \leq -z)$

10.3.1 Calculating Normal Probabilities

Suppose $X \sim \mathrm{N}(100, 9)$. We can calculate the CDF of X using only the standard normal CDF. If $X \sim \mathrm{N}(\mu, \sigma^2)$ then $Z = \frac{X-\mu}{\sigma} \sim \mathrm{N}(0, 1)$.

$$P(X \leq 105) = P(X - \mu \leq 105 - \mu) \tag{61}$$
$$= P\left(\frac{X-\mu}{\sigma} \leq \frac{105-\mu}{\sigma}\right) \tag{62}$$
$$= P\left(Z \leq \frac{105-100}{3}\right) \tag{63}$$
$$= P(Z \leq 1.67) \tag{64}$$
$$= \Phi_Z(1.67) = 0.9525 \tag{65}$$

$$P(99 < X < 103) = P\left(\frac{99-100}{3} \leq \frac{X-\mu}{\sigma} \leq \frac{103-100}{3}\right) \tag{66}$$
$$= P(-1/3 \leq Z \leq 1) \tag{67}$$
$$= P(Z \leq 1) - P(Z \leq -1/3) \tag{68}$$
$$= 0.8413 - 0.3707 = 0.4706 \tag{69}$$

10.3.2 Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ are independent random variables and $E[X_i] = \mu$, $\mathrm{Var}(X_i) = \mu^2$, then if n is large $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately Normally distributed with $E[\bar{X}] = \mu$ and $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$. The distribution of an average tends to be Normal, even when the distribution from which the average is computed is decidedly non-Normal. Generally this works well if $n \geq 30$ (some say 50). Often real world data points are actually weighted averages of a large number of small effects that cannot be measured (e.g. electronic noise, exam grades), which is why so many datasets appear to follow a Normal distribution. The CLT also allows the Normal distribution to be an approximation for many other distributions (Gamma, Chi-Square, Binomial, Poisson, Negative Binomial).

11 Confidence Intervals

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter. The paramater is usually denoted by θ. Often, this parameter is the population mean μ, which is estimated through the

sample mean $\bar{x}$. The **Significance Level** α refers to how "confident" we are that the true value lies within the interval. A typical value for α is 0.05, which corresponds to a 95% confidence interval. Technically speaking, this means that if the confidence intervals are constructed across many separate data analyses of repeated experiments, the proportion of those intervals that contain the true value of the parameter will match the confidence level (typically 95%). We cannot say that there is a 95% probability that the true value lies within the interval, you need Bayesian statistics for that. This is mostly a technicality without much importance, but Statistics professors will beat you over the head with it so it's a good thing to know.

11.1 Known σ

Suppose that somebody weighs themselves on a regular basis and obtains the following data points $\{175, 176, 173, 175, 174, 173, 173, 176, 173, 179\}$. The sample mean $\bar{X} = 174.70$. If he or she knows the standard deviation for the scale is 1.5 pounds, and measurements from the scale follow a Normal distribution, then the sample mean will have a distribution $\mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Therefore,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1) \tag{70}$$

Now, we want to find bounds on $\bar{X}$ that give a 95% confidence interval.

$$0.95 = 1 - \alpha = P(-z \leq Z \leq z) \tag{71}$$

$$= P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \tag{72}$$

$$= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \tag{73}$$

The final equation gives us our 95% confidence interval for μ. In our example, the 95% confidence interval would be $\left[174.7 - 1.96\frac{1.5}{\sqrt{10}}, 174.7 + 1.96\frac{1.5}{\sqrt{10}}\right] = [173.77, 175.63]$. The 1.96 value comes from setting $\Phi(z) = 0.975$ (.025 on each side for a total of .05). The z-value that solves that equation is 1.96, and is known as the critical value (denoted z^*).

11.2 Unknown σ

More realistically, the population standard deviation will not be known. In this case, the procedure is very similar, but the normal distribution is replaced by **Student's t-distribution**. This distribution is the same as the standard Normal distribution except it takes an additional parameter that stretches it out. This parameter is known as the degrees of freedom. As the degrees of freedom approach infinity, the t-distribution converges to the standard Normal. The purpose is to allow for greater uncertainty when the sample size is small and the Central Limit Theorem cannot be applied to assume normality. We now have:

$$P\left(\bar{X} - t^* \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + t^* \frac{\sigma}{\sqrt{n}}\right) \tag{74}$$

The critical value $t*$ can be looked up in tables or retrieved from common statistical software packages.

12 Hypothesis Testing

Hypothesis testing is used to accept or reject statistical hypotheses, which are generally an assumption about a population parameter. The **Null Hypothesis** (H_0) is usually the hypothesis that the sample observations result purely by chance (e.g. the difference between two means is 0). The **Alternative Hypothesis** (H_a) is the hypothesis that the sample observations are influenced by some non-random cause (e.g. the difference between not equal to 0). In hypothesis testing you either reject H_0 (e.g. two means are significantly different from each other) or you fail to reject H_0. The last part is crucial, failing to reject H_0 is not the same as proving H_a. To perform a test, we assume H_0 is true and then see if the data are sufficiently at odds with that assumption. Usually this involves a critical region as defined by a test statistic.

There are two types of error involved: **Type I** and **Type II**. Type I error is rejecting H_0 when H_0 is actually true, and Type II error is failing to reject H_0 when H_a is actually true. The maximum probability of Type I error is known as the significance level α of the test (usually 0.05). The **P-Value** is defined as the smallest α for which H_0 is rejected. Informally, it can be thought of as the probability of an obtaining an outcome as or more extreme than the observed values, relative to H_0. Hypothesis testing is a very general idea that can be covered in entire books, so I'll just get to an example.

A sample of 40 sales recepts from a grcery store has $\bar{x} = 137$ and $\sigma = 30.2$. Use these values to test whether or not the mean sales at the grocery store are different from \$150.

1. Set the null and alternative hypotheses.

$$H_0 : \mu = 150 \tag{75}$$

$$H_a : \mu \neq 150 \tag{76}$$

2. Calculate the test statistic, similar to what we did for confidence intervals.

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{137 - 150}{30.2/\sqrt{(40)}} = -2.2722 \tag{77}$$

3. Set rejection region. This is a two sided hypothesis test, so our critical z-value comes from $\Phi(z) = 0.05/2 = 0.025$.

$$R : |Z| > 2.58 \tag{78}$$

4. Conclude. We see that $|-2.722| = 2.722 > 2.58$, thus our test statistic is in the rejection region. Therefore we reject the null hypothesis and conclude that the mean is significantly different from \$150.

The t-distribution is used instead of the Normal in the same manner as before. In fact, you can perform certain hypothesis tests just by looking at the confidence intervals because the same values are calculated. In the previous example, we know that the true mean μ is significantly different from 170 pounds because that lies well outside of the confidence interval.

13 Conclusion

I've done my best to give an overview of probability and statistics, but of course I have only scratched the surface. Google is pretty good with these topics, and I've tried to put enough information here to allow you to find more complex subjects on your own.