1 Supplementary notes on probability

1.1 The normal distribution

Many naturally occuring variables have distributions that are well-approximated by a "bell-shaped curve", or a normal distribution. These variables have histograms which are approximately symmetric, have a single mode in the center, and tail away to both sides. Two parameters, the mean μ and the standard deviation σ describe a normal distribution completely and allow one to approximate the proportions of observations in any interval by finding corresponding areas under the appropriate normal curve.

In addition, the sampling distributions of important statistics such as the sample mean are approximately normal for moderately large samples.

Characteristics of all normal curves:

- Each bell-shaped normal curve is symmetric and centered at its mean μ.
- The total area under the curve is 1.
- About 68% of the area is within one standard deviation of the mean, about 95% of the area is within two standard deviations of the mean, and almost all (99.7%) of the area is within three standard deviations of the mean.
- The places where the normal curve is steepest are a standard deviation below and above the mean ($\mu - \sigma$ and $\mu + \sigma$).

Standardization: In working with normal curves, the first step in a calculation is invariably to standardize.

$$z = \frac{x - \mu}{\sigma}$$

This z-score tells how many standard deviations an observation x is from the mean. Positive z-scores are greater than the mean, and negative z-scores are below the mean.

If the z-score is known and the value of x is needed, solving the previous equation for x gives

$$x = \mu + z \times \sigma$$

Reading the algebra, this simply states that x is z standard deviations above the mean.

The standard normal distribution: Areas under all normal curves are related. For example, the area to the right of 1.76 standard deviations above the mean is identical for all normal curves. Because of this, we can find an area over an interval for any normal curve by finding the corresponding area under a standard normal curve which has mean $\mu = 0$ and standard deviation $\sigma = 1$.

Using the normal table: The standard normal table is located in an appendix of your textbook. I will use the table that is on the course Web page. You should be familiar with this as well because you will need to use this table for your tests. The table tells you the area to left of a number z (rounded to four decimal places) as z varies from 0 to 4. This is sufficient to find the area under the curve over any interval because all normal curves are symmetric and the total area under any normal curve is one.

You will need to be able to use the table to find areas when the numbers on the axis are known, and to be able to use the table to find numbers on the axis when areas are known.

Always draw a sketch of a normal curve in working out problems.

Example: Find the area to the left of 0.57.

The row labeled 0.5 and the column labeled 0.07 have entry 0.7157. This is the area to the left of 0.57. Because the total area is 1, the area to the right of 0.57 must be $1 - 0.7157 = 0.2843$.

Example: Find the area to the left of -1.43.

Because of symmetry, the area to the left of -1.43 equals the area to the right of 1.43. The table tells us the area to the left of 1.43 is 0.9236. Therefore, the area to the left of -1.43 is $1-0.9236 = 0.0764$.

Example: Find the area between -1.37 and 2.33.

The area between -1.37 and 2.33 is the area to the left of 2.33 minus the area to the left of -1.37. The area to the left of 2.33 is 0.9901. The area to the left of -1.37 is the area to the right of 1.37 which is $1 - 0.9147 = 0.0853$. The desired area is $0.9901 - 0.0853 = 0.9048$.

Example: Find the value z so that the area to the right of z is 0.6000.

The area to the left of z is 0.4000 and z is negative, so z will be the negative of the number where the area to the left is 0.6000. The value $z = 0.25$ corresponds to an area 0.5987, which is as close as possible. Therefore, $z = -0.25$.

Examples Ridge counts in finger prints are approximately normally distributed with a mean $\mu = 140$ and standard deviation $\sigma = 50$.

What proportion of individuals have: (a) a ridge count more than 200, and (b) a ridge count less than 100. (c) Find the 90th percentile of this variable.

Solution to (a): The area to the right of 200 under the given normal curve is equal to the area to the right of $z = (200 - 140)/50 = 1.20$. under the standard normal curve. The area to the left of 1.20 is 0.8849 from the table, so the answer is $1 - 0.8849 = 0.1151$.

Solution to (b): The area to the left of 100 under the given normal curve is equal to the area to the left of $z = (100 - 140)/50 = -0.80$ under the standard normal curve. Looking up $z = 0.80$ gives 0.7881, The area we want is $1 - 0.7881 = 0.2119$.

Solution to (c): The 90th percentile is above the mean at a point x. According to the normal table, the z-score of x must be very close to 1.28, so $x = 140 + 1.28(50) = 204$, a number slightly higher than 200, as was expected after doing part (a).

1.2 The binomial distribution

The binomial distribution arises from counting the number of heads in a prespecified number of coin tosses. This is a model for the way that data is produced for a vast number of examples in statistics. In particular, we will use this model when examining the proportion of a random sample that belongs to a particular category.

Every binomial random variable is described by two parameters: n is the number of trials and p is the probability that an individual trial is a success.

The binomial setting: A trial has one of two possible values. One is called a "success" and the other is called a "failure". We want to count the number of successes.

The binomial distribution is appropriate when we have this setting:

1. there are a fixed number of trials;
2. there are two possible outcomes for each trial;
3. the trials are independent of one another;
4. there is the same chance of success for each trial;
5. we count the number of successes

The textbook gives examples of how to calculate the probability of individual outcomes.

The normal approximation to the binomial distribution: Consider this example: find the probability that there are 410 or more successes in 500 independent trials when the probability of success on a single trial is 0.8. An exact expression of this probability is

$$P(X \geq 410) = \sum_{x=410}^{500} \frac{500!}{x!(500-x)!}(.75)^x(.25)^{500-x}$$

Even with a calculator, this is an intimidating computation. Statistical software gives the answer numerically as 0.1437.

When the number of trials is sufficiently large, a graph of the binomial distribution resembles the familiar shape of the normal distribution. When the number of expected successes and failures in the trials are each sufficiently large, an area under the normal curve will be a good numerical approximation to the exact binomial computation. A good rule of thumb is that if $np \geq 5$ and $n(1-p) \geq 5$, then an approximation will be good (usually to about two or three digits).

There are many normal distributions, each described completely by μ and σ. To find the one which matches a binomial distribution best, let $\mu = np$ be the mean (or expected value) and $\sigma = \sqrt{np(1-p)}$ be the standard deviation. This μ will be the balancing point of the binomial distribution. You expect the random variable to be close to μ, but it may not be μ exactly. In fact, μ may not even be an integer and may not be a possible value of the random variable. Although the random variable could take on any integer value between 0 and n, it is highly unlikely to be more than a two or three standard deviations σ from μ. You may interpret σ as the rough size of a typical deviation from the mean.

Follow these steps to apply the approximation.

1. Check to see if n is large enough for the given p.
2. Write the outcome in the form $a \leq X \leq b$ where X is the random number of successes in the n trials.
3. Find μ and σ.
4. Find the area between $a - 0.5$ and $b + 0.5$ under a normal curve with mean μ and standard deviation σ.

For the problem above, $500(0.8) \geq 5$ and $500(0.2) \geq 5$ so the sample size is easily large enough to get a good approximation. We want to find $P(410 \leq X \leq 500)$ with $n = 500$ and $p = 0.8$. $\mu = 500(0.8) = 400$ and $\sigma = \sqrt{500(0.8)(0.2)} = 8.94$. The z-score of 409.5 is $(409.5 - 400)/8.94 = 1.06$. The z-score of 500.5 is $(500.5 - 400)/8.94 = 11.2$. The area between 1.06 and 11.2 under the standard normal curve is $1 - 0.8554 = 0.1446$, which is close to the exact 0.1437.

1.3 The Poisson distribution

The textbook claims that the Poisson distribution is a good approximation of the binomial distribution when n is very large and p is very small. We will use the Poisson distribution to describe the probability of counting the number of rare events. It depends on a single parameter λ and satisfies

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!}$$

for x a nonnegative integer. When n is very large and p is very small, the binomial probabilities are numerically very close to the Poisson probabilities with $\mu = np$.

For example, say that over the past several years, there have been an average of 13 new cases of cancer of the esophogus diagnosed in Pittsburgh. The probability that any individual gets a this type of cancer is rare. The population of Pittsburgh is fairly stable so it is not too outrageous to make the approximation that it stays constant. The population of the Pittsburgh metropolitan area is the large number n and the probability an individual is diagnosed with this form of cancer is the small number p. We may expect that the number of diagnoses in the upcoming year is a random Poisson variable with $\mu = 13$. If so, the probability that there are exactly 10 new cases would be

$$P(X = 10) = \frac{e^{-13}(13)^{10}}{10!} = 0.086$$

This is far more reasonable to compute than plugging into the binomial distribution formula with $n = 2,500,000$ and $p = 13/n$.