STAT 2120, Fall 2012: Notes on Topic 5

Conditional probability:

- A conditional probability, $P(B|A)$, gives the probability of some event, B, under the condition that some other event, A, has definitely occurred.
- The general multiplication rule is $P(A \text{ and } B) = P(A)P(B|A)$.
  - This rule extends to multiple events as $P(A \text{ and } B \text{ and } C) = P(A)P(B|A)P(C|A \text{ and } B)$, *etc.*
  - Rearranging this rule yields $P(B|A) = P(A \text{ and } B)/P(A)$, which serves as a definition of conditional probability.
  - If A and B are independent, then $P(B|A) = P(B)$, which is quickly derived from the previous property $P(A \text{ and } B) = P(A)P(B)$.
- A tree diagram is convenient way of organizing one's thinking when working with conditional probabilities.
  - Each branch extending from some event to one of possibly many other events represents a segment of a possible path through the stages of a problem. It is labeled by the conditional probability of the latter event given the former.
  - The conditional probabilities associated with all of the branches extending from the same event must sum to one.
  - Each complete path from the first to the last stage of a problem represents the overlap of the events along that path. Its probability is calculated by multiplying the corresponding conditional probabilities.
  - The probability of some event at the final stage of the problem is calculated by adding the product of probabilities along all complete paths that lead to that event.
- Bayes's rule is: $P(A|B) = P(B|A)P(A)/\{P(B|A)P(A) + P(B|A^c)P(A^c)\}$.
  - It is often easier work with the definition of conditional probability, while manipulating probabilities in a tree diagram, than work with the formula for Bayes's rule.
  - The numerator of Bayes's rule reflects the multiplication of conditional probabilities along a tree diagram's complete path through A and then B.
  - The denominator of Bayes's rule reflects the addition of probabilities of all the complete paths in a tree diagram that lead to event B.

Binomial distributions:

- The binomial distributions provide a theoretical model for count data having a fixed maximum.
- The binomial setting is defined as follows.
  - A fixed number, n, of trials (*i.e.*, chance happenings) are observed.
  - The trials are independent. That is, knowing the outcome of any one trial will not affect the probabilities governing any other trial.
  - Each trial has the same two possible outcomes, whose generic labels are S (for "success") and F (for "failure").
  - The success probability, $p = P(S)$, is the same for each trial.
- Some properties of the binomial setting are:
  - The sample space consists of 2^n possible outcomes, corresponding to the number of possible length-n sequences of S and F.
  - Each possible outcome has probability $p^{\#S}(1-p)^{\#F}$, where $\#S$ and $\#F$ count the respective number of S and F in n trials.
  - There are $\binom{n}{k}$ possible outcomes with S appearing exactly k times.
- The random variable, X, that counts the number of S in the binomial setting is called a binomial random variable and is said to have a binomial distribution.
- Some properties of the probability model for a binomial random variable, X, are:
  - The sample space consists of $n+1$ possible outcomes, $S = \{0, 1, \ldots, n\}$.
  - Probabilities are assigned as $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$. These are sometimes called binomial probabilities.
  - The mean and standard deviation are $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$.
- Some approaches for finding binomial probabilities are:
  - Use the formula $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.
  - Use a binomial table, such as that on p. 325.
  - Use the Excel function =binomdist($k, n, p, 0$) for $P(X = k)$ or =binomdist($k, n, p, 1$) for $P(X \leq k)$. The latter are called cumulative probabilities.
  - Use a Normal approximation, $P(X \leq k) \approx P\left(Z \leq (k - np)/\sqrt{np(1-p)}\right)$. A rule of thumb is to apply the Normal approximation when $np \geq 10$ and $n(1-p) \geq 10$.

Poisson distributions:

- The Poisson distributions provide a theoretical model for open-ended counts.
- The Poisson setting is defined as follows:
  - "Success points" are counted within a fixed region or time-interval, *etc.*, which is a continuum and may be subdivided into arbitrarily small "units of measure."
  - Counts of success points are independent between any nonoverlapping units of measure.
  - The mean count of success points in any unit of measure is proportional to its size.
  - The probability of two or more success points in the same unit of measure becomes arbitrarily small as the size of the unit shrinks.
- The random variable, X, that counts the number of success points in the Poisson setting is called a Poisson random variable and said to have a Poisson distribution.
- Some properties of the probability model for a Poisson random variable, X, are:
  - The sample space is $S = \{0, 1, \dots\}$, which is infinite, but may still be counted; thus, X is a discrete random variable.
  - Probabilities are assigned as $P(X = k) = e^{-\mu}\mu^k/k!$, where μ is the mean count of success points and $e \approx 2.71828$ is the base of the natural logarithms. These are sometimes called Poisson probabilities.
  - The mean and standard deviation are $\mu_X = \mu$ and $\sigma_X = \sqrt{\mu}$.
- Some approaches for finding Poisson probabilities are:
  - Use the formula $P(X = k) = e^{-\mu}\mu^k/k!$.
  - Use the Excel function =poisson($k, \mu, 0$) for $P(X = k)$ or =poisson($k, \mu, 1$) for cumulative probabilities $P(X \leq k)$.
- Some points to consider when pondering the Poisson distribution as a model for data:
  - If X and Y are Poisson random variables counting the success points in nonoverlapping regions or time-intervals, *etc.*, then $Z = X + Y$ is a Poisson random variable with mean $\mu_Z = \mu_X + \mu_Y$.
  - If μ is the mean count of success points per unit of space or time, then $\alpha\mu$ is the mean count of success points in a region or time-interval α units in size.

STAT 2120, Fall 2012: Notes on Topic 6

Introduction:

- Probability calculations help distinguish patterns seen in data between those that are due to chance and those that reflect a real feature of the phenomenon under study.
- The two most prominent types of formal statistical inference are confidence intervals and tests of significance.
  - These report probabilities describing what would happen in the "long run" if the experiment was repeated many, many times.
  - The probabilities derive from sampling distributions, based on a probability model.
- Concepts will be introduced for inference on the mean, μ, of a population.
  - The data are taken to derive from a sample of size n, on which $\bar{x}$ and s are calculated.
  - The population variance σ will be (unrealistically) treated as known. Substitute s for σ until we learn how to work when σ is unknown.
  - A population refers to the entire collection of items from which a sample is drawn.

Estimating with "confidence:"

- The sample mean, $\bar{x}$, provides an unbiased estimate of μ. A confidence interval is intended to provide such an estimate with an indication of its variability.
- The basic reasoning underlying a confidence interval for μ is as follows:
  - By the central limit theorem, $\bar{x}$ is approximately $N(\mu, \sigma/\sqrt{n})$. Thus, it is natural to state distances of $\bar{x}$ from μ in units of (the sample mean's) standard deviation, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.
  - The probability of $\bar{x}$ lying within a certain distance, $a\sigma_{\bar{x}}$, from μ, $C = P(-a\sigma_{\bar{x}} \leq \bar{x} - \mu \leq a\sigma_{\bar{x}})$, provides a degree of confidence in such an assessment. The quantity C is called the confidence level.
  - One would state being $C100\%$ confident of μ lying between the bounds $\bar{x} \pm a\sigma_{\bar{x}}$. A confidence interval refers to the interval between those bounds.
  - The quantity $a\sigma_{\bar{x}}$ is a margin of error. A smaller margin of error indicates a more precise inference.
  - In the "long run," after repeated experimentation, $C100\%$ of confidence intervals would cover μ.
- Given a desired confidence level, C, a general formula for a confidence interval for μ is $\bar{x} \pm z^* \sigma/\sqrt{n}$.
  - The quantity z^* is such that $C = P(-z^* \leq Z \leq z^*)$, where Z is $N(0,1)$. It is called a critical value of the standard normal distribution.
  - The associated margin of error is $m = z^* \sigma/\sqrt{n}$.
  - The confidence level, C, is exactly correct if the population has a Normal distribution, and approximately correct when n is large.
- Elementary behavior of confidence intervals, in terms of the margin of error, $m = z^* \sigma/\sqrt{n}$:
  - Requiring higher confidence, by increasing C, increases the m.
  - A smaller population variance, σ, decreases m.
  - A larger sample size, n, decreases m.
- When planning a sample, the sample size may be chosen to target a desired m from a desired C.
  - The relevant sample size formula is $n = (z^* \sigma/m)^2$.
- The confidence interval formulas of this section are valid only in specific circumstances.
  - The sample must have been drawn by SRS.
  - Since $\bar{x}$ is not resistant, neither is the associated confidence interval.
  - Derivation of the confidence interval formula relies on the central limit theorem. If the population is possibly non-Normal, $n \geq 15$ is usually sufficient for accuracy, absent extreme outliers or strong skewness.
  - Knowledge of σ has been assumed, but the formula $\bar{x} \pm z^* s/\sqrt{n}$ is valid for large samples.
  - The margin of error accommodates sampling variability, but not errors due to sampling bias, such as undercoverage and nonresponse.

Tests of significance:

- A significance test aims to assess the truth of a hypothesis through comparison with observed data.
  - A hypothesis is a statement about the parameters of a population or model.
  - The null hypothesis, H_0, states the status quo (*e.g.*, no effect of a new treatment).
  - The alternative hypothesis, H_a, states our suspicion of what is true (*e.g.*, the new treatment is effective).