STAT 50, A Review of Formulas for Test 1

A. Summary Measures of Samples and Finite Populations.

Sample mean. $\bar{x} = (\sum x)/n$, where n is the sample size and the sum is over all values in the sample.

Sample standard deviation. $s_x = \sqrt{\frac{1}{n-1}\sum(x-\bar{x})^2}$ (defining formula)

$$s_x = \sqrt{\frac{\sum x^2 - (\sum x)^2/n}{n-1}} \text{ (computing formula)}$$

Population mean. $\mu = (\sum x)/N$, where N is the population size and the sum is over all values in the population.

Population standard deviation. $\sigma_x = \sqrt{\frac{1}{N}\sum(x-\mu)^2}$ (defining formula)

$$\sigma_x = \sqrt{\frac{\sum x^2 - (\sum x)^2/N}{N}} \text{ (computing formula)}$$

Z-scores: For a population data set with mean μ and standard deviation σ, the z-score of a value x is $z = \frac{x-\mu}{\sigma}$. It is similar for sample data: $z = \frac{x-\bar{x}}{s}$.

Five number summary. Min, Q_1, Q_2 (median), Q_3 and max.

IQR and limits. $IQR = Q_3 - Q_1$, LL $= Q_1 - 1.5(IQR)$ and UL $= Q_3 + 1.5(IQR)$.

Linear functions. If $y = ax + b$, then $\bar{y} = a\bar{x} + b$ and $s_y = |a|\, s_y$. Similar results hold for population parameters (including the mean and standard deviation of a random variable):

$\mu_y = a\mu_x + b$, and $\sigma_y = |a|\,\sigma_x$.

B. Probability, Conditional Probabilities and Independence.

- For any event A, $P(A) + P(A^c) = 1$ (the complement rule).
- If $A \cap B = \emptyset$ (if A and B are mutually exclusive),then $P(A \cup B) = P(A) + P(B)$. This special addition rule is also valid for any number of pairwise disjoint events $A_1, A_2..., A_n$:

$$P(A_1 \cup A_2 \cup ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n).$$

- General addition rule for two events. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- Conditional probability. $P(A|B) = \frac{P(A\cap B)}{P(B)}$, if $P(B) > 0$.
- Multiplication rule. For any two events A and B, $P(A \cap B) = P(B)P(A|B)$, with $P(B) > 0$.
- Total probability formula. If $\{A_1, A_2..., A_n\}$ is a partition of a sample space and B is an event, then

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + ... + P(B \cap A_n) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + ... + P(B|A_n)P(A_n). \\ &= \sum_{j=1}^{n} P(B|A_j)P(A_j). \end{aligned}$$

- Bayes' Rule. If $\{A_1, A_2..., A_n\}$ is a partition of the sample space and $P(B) > 0$, then the *posterior* probabilities $P(A_k|B)$, for each $k = 1, 2, ...n$, are given by the Bayes' rule:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}.$$

- Two events A and B are **independent** if and only if any one of the followings holds:
  (1) $P(A|B) = P(A)$
  (2) $P(B|A) = P(B)$
  (3) $P(A \cap B) = P(A)P(B)$ - this is also called the special multiplication rule

- The (**mutual**) independence of multiple events. We need to verify that the special multiplication rule holds for all possible intersections of different combinations of these events. For three events A, B and C, all of the following identities should hold for the collection {A, B, C} to be (mutually) independent:

$$\begin{aligned} P(A \cap B) &= P(A)P(B), P(A \cap C) = P(A)P(C), P(C \cap B) = P(C)P(B), \\ P(A \cap B \cap C) &= P(A)P(B)P(C). \end{aligned}$$

- If n events $A_1, A_2..., A_n$ are mutually independent, then $P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)P(A_2)...P(A_n)$.

C. Counting.

By applying the multiplication rule of counting, we obtain the *permutation* formula for total number of ordered selections of k items from a list of n items without replacement:

- $P(n, k) = n(n-1)(n-2)...(n-k+1)$ or $P(n, k) = \frac{n!}{(n-k)!}$ in compact form.
- Combination formula (for unordered selections without replacement): $C(n, k) = \frac{n!}{k!(n-k)!}$.
- Another notation for $C(n, k)$ is $\binom{n}{k}$.

D. Random Variables.

For a discrete random variable (r.v.) X, let $p(x) = P(X = x)$ be the probability mass function (pmf).

- Then the **mean** (or expected value) of X is given by $E(X) = \mu = \sum_x xp(x)$.
- If X and Y are two random variables with means μ_X and μ_Y, respectively, then the mean of $X+Y$ is $\mu_X + \mu_Y$.
- The **variance** of X is:

$$\begin{aligned} \sigma^2 &= \sum_x (x-\mu)^2 p(x) \text{ (defining formula)} \\ &= \sum_x x^2 p(x) - \mu^2 \text{ (computing formula).} \end{aligned}$$

- The standard deviation, σ, is the square root of the variance.
- *The cumulative distribution function (cdf)*: $F(x) = P(X \leq x)$.
- For a discrete r.v. with pmf $p(x)$, $F(x) = \sum_{t \leq x} p(t)$, which is a nondecreasing and right-continuous function.
- For a continuous r.v. X with a probability density function (pdf) $f(x)$, the cdf is is

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

which is a continuous and non-decreasing function. Moreover, F is differentiable where f is continuous and satisfies $F'(x) = f(x)$ at such points of continuity.
- For probabilities of such a continuous r.v. with values in an interval, we can use

$$P(a < X < b) = \int_{a}^{b} f(t)dt = F(b) - F(a).$$

- If m is the median of X, then m satisfies $F(m) = 0.5$. Similarly, if x_p is the p^{th} percentile of X, then it satisfies the relation $F(x_p) = p/100$. For example, if y is 90^{th} percentile of X, then $F(y) = 0.9$.
- The mean and variance of a continuous r.v. are based on some integral-based formulas (but they are not included in the upcoming test). For example, the mean is given by $\mu = \int_{-\infty}^{\infty} xf(x)dx$, if it is a finite number. If the pdf is zero in an interval, then we ignore that part in integral computations.