ST2334 Final Examination Cheat-sheet

Part 1 Statistical Description

1. Random variable: categorical (nominal, ordinal) & numerical (discrete, continuous).
2. Sampling: *non-probability* (quota, convenience, judgment) & *probability* (simple random, systematic, cluster – units in a cluster are like the population, stratified – units in a stratum are homogeneous).
3. To display categorical variables: pie chart, bar chart, pareto chart.
4. To display numerical variables: histogram, stem-and-leaf plot, dot-plot.
5. Measures of location: mean, median, mode (unimodal, bimodal, trimodal) & percentile (quantile).
6. Measures of spread: variance (standard deviation), coefficient of variation (CV), range, inter-quartile range (IQR), box plot (box-and-whisker plot), five-number summary.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

7. Shape of a distribution: symmetric & skewed (left, right). When left skewed, mean < median < mode; right skewed, mode < median < mean.

Part 2 Probability

1. Terms: experiment, outcome, sample space, event.
2. Event: null, union, intersection, complement, mutually exclusive, independent.
3. Interpretation of probability: equally-likely outcomes, frequency interpretation, personal probability.
4. Conditional probability: $P(B|A) = \frac{P(A\cap B)}{P(A)}$
5. Two events are independent if and only if: 1) $P(A\cap B) = P(A)P(B)$; 2) $P(A|B) = P(A)$; 3) $P(B|A) = P(B)$.
6. Partition: If $B_1, B_2, \dots, B_n$ are mutually exclusive and $B_1 \cup B_2 \cup \dots \cup B_n = S$, we call them being a partition of S.
7. Total probability: $P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i)$ for any partition of S.
8. Bayes' Theorem: $P(B_k|A) = \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)}$.
9. Two *non-trivial mutually-exclusive* events must be dependent.

Part 3 Discrete Random Variable

1. Probability mass function (PMF): 1) $f(x_i) \geq 0$; 2) $\sum f(x_i) = 1$.
2. Variance: 1) $Var(a+bX) = b^2 \cdot Var(X)$; 2) $Var(X) = E(X^2) - (E(X))^2$.
3. Bernoulli distribution: If $W \sim B(n,p)$, then $P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$, we also have $E(x) = np$ & $Var(x) = np(1-p)$.
4. Geometric distribution: If $X \sim Geom(p)$, then $P(X=k) = (1-p)^{k-1}p$, we also have $E(x) = \frac{1}{p}$ & $Var(x) = \frac{1-p}{p^2}$, with CDF being $F(x) = 1-(1-p)^x$.
4. Poisson distribution: If $X \sim Pois(\lambda)$, then $P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}$, we also have $E(x) = \lambda$ & $Var(x) = \lambda$.
5. When $n \geq 20, p \leq 0.05$ or $n \geq 100, np \leq 10$, we can use Poisson to approximate Binomial by $B(n, \lambda/n) \approx Pois(\lambda)$.
6. Cumulative distribution function (CDF) can be applied to discrete & continuous random variables, it is non-decreasing $F(x) = P(X \leq x)$.

Part 4 Continuous Random Variable

1. Probability density function (PDF): 1) $f(x_i) \geq 0$; 2) 1 $\int_{-\infty}^{\infty} f(x)\,dx = 1$.
2. Mean: $\mu = E(X) = \int_{-\infty}^{\infty} x f_x(x)\,dx$ and $E\big(g(x)\big) = \int_{-\infty}^{\infty} g(x) f_x(x)\,dx$.
3. Normal distribution: If $X \sim N(\mu, \sigma^2)$, then $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/\sigma^2}$, we also have $E(X) = \mu$ and $var(X) = \sigma^2$.
4. When np and $n(1-p)$ are both larger than 15, we can use normal distribution to approximate Binomial distribution 1 $X \sim N(np, np(1-p))$.
5. When approximating a discrete random variable using a continuous random variable, *continuity correction* is necessary.
6. Uniform distribution: If $X \sim U(a,b)$, then 1 $f_X(x) = 1/(b-a)$, we also have $E(x) = (a+b)/2$, $var(x) = (b-a)^2/12$ and $F(x) = (x-a)/(b-a)$.
7. Exponential distribution: If $X \sim Exp(\lambda)$, then $f_X(x) = \lambda e^{-\lambda x}$, we also have $E(x) = 1/\lambda$, $var(x) = 1/\lambda^2$ and $F_X(x) = 1 - e^{-\lambda x}$. It is also memory-less since $P(X > s+t \mid X > s) = P(X+t)$.
8. Chebyshev's inequality: $P(|X-\mu| \geq k\sigma) \leq 1/k^2$.

Part 5 Joint Distribution

1. X and Y are independent if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x and y.

2. Marginal distribution: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$.

3. Conditional distribution: $f_{X\mid Y}(x \mid y) = f_{X,Y}(x,y)/f_Y(y)$.

4. Covariance: $cov(X,Y) = E(XY) - E(X)E(Y)$. If X and Y are independent, $cov(X,Y) = 0$. Also, $var(aX + bY) = a^2 var(X) + b^2 var(Y) + 2ab \cdot cov(X,Y)$.

Part 6 Sampling & Sampling Inference

1. For random samples of size n taken from population with mean μ and variance σ^2, the sample mean has $E(\bar{X}) = \mu$ and $var(\bar{X}) = var(X)/n$.

2. Law of large numbers (LLN): $\lim_{n\to\infty} P(|\bar{X} - \mu| > \epsilon) = 0$ for any $\epsilon \in \mathbb{R}$.

3. Central limit theorem (CLT): $\lim_{n\to\infty} \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. It approximates well the sampling distribution when $n \geq 30$.

4. When X_i is normal, $\bar{X} \sim t_{n-1}$ for small n.

5. When X_i is normal, $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ with parameter $v = n - 1$.

6. Maximum error: With probability $1 - \alpha$, the error $|\bar{X} - \mu|$ is within $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

7. Four cases of point estimation:

Case	σ	n	Population	Statistic	E	n needed for desired E and α
I	known	any	Normal	$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2$
II	known	large	any	$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2$
III	un-known	small	Normal	$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$	$t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}$	$\left(\frac{t_{n-1;\alpha/2}\,S}{E}\right)^2$
IV	un-known	large	any	$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$	$t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}$	$\left(\frac{t_{n-1;\alpha/2}\,S}{E}\right)^2$

8. If $P(A < \mu < B) = \alpha$, then (A, B) is the α confidence interval for μ.

9. Five steps to hypothesis testing:

1) Set up null vs alternative hypotheses.
2) Determine the level of significance.
3) Identify statistic, distribution, small/large sample & rejection criteria.
4) Compute based on data and compare by one/double-side.
5) Make conclusion on whether to reject num hypothesis.

10. Two types of errors:

	Not reject H_0	**Reject H_0**
H_0 is true	Correct Decision	**Type I error**
H_0 is false	**Type II error**	Correct Decision

11. Usually, we can also use *p-value*, the "observed" one/double-sided confidence interval.

Part 6 Sampling Inference of Two Means

1. There are usually two types of designs for comparing two treatments: independent complete randomization & matched pair randomization.

2. For two independent large samples, $(\bar{X} - \bar{Y}) \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$.

3. For two small samples which are both normally distributed, the equal variance assumption holds if and only if $0.5 \leq s_1/s_2 \leq 2$.

4. We define the pooled estimator as $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$.

5. Under equal variance assumption, $T = \frac{(\bar{X}-\bar{Y})-\delta}{S_p\sqrt{1/n_1+1/n_2}} \sim t_{n_1+n_2-2}$.

6. Without equal variance assumption, for two small samples which are both normally distributed, $T = \frac{(\bar{X}-\bar{Y})-\delta}{\sqrt{S_1^2/n_1+S_2^2/n_2}} \sim t_k$ where k is defined as

$$k = \left\lfloor \frac{(S_1^2/n_1+S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}} \right\rfloor$$

7. For matched pair samples, the difference of two means can be considered as single-variable sampling which follows normal distribution (for large samples) or t-distribution (for small samples under normal).
