ST2334 Cheatsheet AY22/23 —— @JasonYapzx

Chapter 1

Observation: Recording of information, numerical or categorical
Statistical Exp: Procedure that generates a set of observations
Sample Space: Set of all possible outcomes of a statistical experiment, represented by the symbol S.
Sample Points: Every outcome in a sample space.
Events: Subset of sample space.
Simple/Compound Event: Exactly one/More than one outcome or sample point.
Sure/null Event: Sample space/Event with no outcomes.

Operations on Events

- **Complement**: A' Elements not in A
- **Mutually Exclusive:** $A \cap B = \emptyset$
- **Union:** $A \cup B$ contains A or B or both elements
- **Intersection:** $A \cap B$ contains elements common to both

1. $A \cap A' = \emptyset$
2. $A \cap \emptyset = \emptyset$
3. $A \cup A' = S$
4. $(A')' = A$
5. $(A \cap B)' = A' \cup B'$
6. $(A \cup B)' = A' \cap B'$
7. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
8. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
9. $A \cup B = A \cup (B \cap A')$
10. $A = (A \cap B) \cup (A \cap B')$

$A \subset B$ if all elements in event A are in event B, if $A \subset B$ and $B \subset A$ then $A = B$. We assume **contained** means proper subset.

Permutations and Combinations

- P: Arrange r objects from n objects where $r \leq n$, ${}_nP_r = \frac{n!}{(n-r)!}$
- P: Number of ways around a circle $= (n-1)!$
- C: No. of ways to select r from n without regard to order: $\binom{n}{r} = {}_nC_r = \frac{n!}{r!(n-r)!}$
- **Conditional:** $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$, if $\Pr(A) \neq 0$
- **Multiplicative:** $\Pr(A \cap B) = \Pr(A)\Pr(B|A) = \Pr(B)\Pr(A|B)$
- **LoTP:** $\Pr(B) = \sum_{i=1}^{n} \Pr(B \cap A_i) = \sum_{i=1}^{n} \Pr(A_i)\Pr(B|A_i)$
- **Bayes:** $\Pr(A_k|B) = \frac{\Pr(A_k)\Pr(B|A_k)}{\sum_{i=1}^{n} \Pr(A_i)\Pr(B|A_i)}$

Properties of Independent Events

1. $\Pr(B|A) = \Pr(B)$ and $\Pr(A|B) = \Pr(A)$
2. A and B cannot be mutually exclusive if they are independent, supposing $\Pr(A), \Pr(B) > 0$
3. A and B cannot be independent if they are mutually exclusive
4. Sample space S and empty set $\emptyset$ are independent of any event
5. If $A \subset B$, then A and B are dependent unless $B = S$.

Birthday Problem $p_n = \Pr(A) = 1 - q_n$, once have 23 people, probability exceeds 0.5. How large does a group of randomly selected people have to be such that the probability that someone is sharing his or her birthday with me is larger than 0.5? n s.t. $1 - (\frac{364}{365})^n \geq 0.5$. $1 - \frac{n \times n-1 \times \cdots \times n-p+1}{n^p}$, n is the number of days, p is number of people.

Chapter 2

Random Variable: $R_X = \{x | x = X(s), sS\}$
Equivalent: If $A = \{s \in S | X(s) \in B\}$, A and B are equivalent
Probability Function: 1. $f(x_i)0$ for all x_i 2. $\sum_{i=1}^{\infty} f(x_i) = 1$
Continuous Random Variable: The probability density function (p.d.f.) $f(x)$ of a continuous random variable must satisfy the following conditions

- $f(x)0$ for all $x \in R_X$. This also means that we may set $f(x) = 0$ for $x \notin R_X$, i.e. $\Pr(A) = 0$ does not imply $A =$
- $\int_{R_X} f(x)dx = 1$, $\int^{\infty} \infty f(x)dx = 1 \rightarrow f(x) = 0$ for x not in R_X
- For any $(c, d) \subset R_X, c < d$, $\Pr(cXd) = \int_c^d f(x)dx$
- $\Pr(X = x_0) = \int_{x_0}^{x_0} f(x)dx = 0$

Cumulative Distribution Function c.d.f
$F(x)$ as cumulative distribution function (c.d.f.) of the random variable X where $F(x) = \Pr(Xx)$. For any ab, $\Pr(aBb) = \Pr(Xb) - \Pr(X < a) = F(b) - F(a^-)$ where a^- is the largest possible value of X that is strictly less than a

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

Expectation

- **Discrete:** $\mu_X = E(X) = \sum_i x_i f_X(x_i) = \sum_x x f_X(x)$
- **Cont.:** $\mu_X = E(X) = \int_{-\infty}^{\infty} x f_X(x)dx$; $E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$

Variance $g(x) = (x - \mu_X)^2$, leads us to the definition of variance.
$\sigma_X^2 = V(X) = E[(X - \mu_X)^2] =$
$$\begin{cases} \sum_x (x - \mu_X)^2 f_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)dx, & \text{if } X \text{ is continuous.} \end{cases}$$

- $\sigma_X = \sqrt{V(X)}$
- $V(X) = E(X^2) - [E(X)]^2$
- $V(X) = E(X^2) - [E(X)]^2$
- $V(aX + b) = a^2V(X)$
- $E(aX + b) = aE(X) + b$, where a and b are constants
- $V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X, Y)$, if X, Y independent, $Cov(X, Y) = 0$

Chapter 3

(X, Y) is a two-dimensional random variable, where X, Y are functions assigning a real number to each $s \in S$.
Range Space: $R_{X,Y} = \{(x, y) | x = X(s), y = Y(s), sS\}$

Joint Probability — Discrete Random Variables
With each possible value (x_i, y_j), we associate a number $f_{X,Y}(x_i, y_j)$ representing $\Pr(X = x_i, Y = y_j)$ and satisfying the following conditions:

1. $f_{X,Y}(x_i, y_j)0$ for all $(x_i, y_j) \in R_{X,Y}$.
2. $\sum^{\infty} i = 1 \sum^{\infty} j = 1 f_{X,Y}(x_i, y_j) = \sum^{\infty} i = 1 \sum^{\infty} j = 1 \Pr(X = x_i, Y = y_j) = 1$.

Joint Probability — Continuous Random Variables
$f_{X,Y}(x, y)$ is called a **joint probability density function** if it satisfies the following:

1. $f_{X,Y}(x, y)0$ for all $(x, y) \in R_{X,Y}$.
2. $\iint_{(x,y) \in R_{X,Y}} f_{X,Y}(x, y)dxdy = 1$ or $\int_{\infty}^{\infty} \int_{\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1$.

Marginal Probabilty Densities

- **Discrete:** $f_X(x) = \sum_y f_{X,Y}(x, y)$ and $f_Y(y) = \sum_x f_{X,Y}(x, y)$
- **Cont:** $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$, $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx$

Conditional Probabilty Densities
The conditional distribution of Y given that $X = x$ is given by $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$, if $f_X(x) > 0$, for each x within the range of X. Flip the variables for X given $Y = y$.

Independent Random Variables
Random variables X and Y are independent if and only if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all x, y, extendable to n variables.

Expectation and Covariance

1. Discrete, $E(g(X, Y)) = \sum_x \sum_y g(x, y)f_{X,Y}(x, y)$
2. Continous $E(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f_{X,Y}(x, y)dydx$
3. $E(XY) = \int \int xy(f(x, y))dydx$
4. $Cov(X, Y) = E(XY) - \mu_X\mu_Y$.
5. If X and Y are independent, then $Cov(X, Y) = 0$. However, $Cov(X, Y) = 0$ does not imply independence.
6. $Cov(aX + b, cY + d) = acCov(X, Y)$.
7. $V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X, Y)$.
8. If $X \perp Y$, $Var(aX - bY) = a^2V(X) + b^2V(Y)$ $(Cov = 0)$
9. $Var(X - Y) = V(X) + V(Y)$

Chapter 4

Discrete Uniform Distributions
If the random variable X assumes the values $x_1, x_2, \cdots, x_k$ with equal probability, then the random variable X is said to have a discrete uniform distribution and the probability function is given by $f_X(x) = \frac{1}{k'}$, $x = x_1, x_2, \cdots, x_k$, and 0 otherwise.

1. Mean, $\mu = E(X) = \sum_{i=1}^{k} x_i \frac{1}{k} = \frac{1}{k}\sum_{i=1}^{k} x_i$.
2. Variance, $\sigma^2 = V(X) = \sum_{\text{all } x}(x - \mu)^2 f_X(x) = \frac{1}{k}\sum_{i=1}^{k}(x_i - \mu)^2$.
3. Variance, $\sigma^2 = E(X^2) - \mu^2 = \frac{1}{k}(\sum_{i=1}^{k} x_i^2) - \mu^2$.

Bernoulli Distributions
Bernoulli experiments only have two possible outcomes, and we can code them as 1 and 0. A random variable X is defined to have a Bernoulli distribution is the probability function of X is given by $f_X(x) = p^x(1 - p)^{1-x}, x = 0, 1$, where $0 < p < 1$. $f_X(x) = 0$ for all other X values.

1. $(1 - p)$ is often denoted by q.
2. $\Pr(X = 1) = p$ and $\Pr(X = 0) = 1 - p = q$.
3. Mean, $\mu = E(X) = p$ — Variance, $\sigma^2 = V(X) = p(1 - p) = pq$.

Discrete Distributions — Formula Sheet

- **Binomial** $\sim B(n, p)$**:** Bernoulli is speical case when $n = 1$ — number of successes in n trials
- **Negative Binomial** $\sim NB(k, p)$**:** Consider a binomial experiment, except that trials *repeated* until a fixed number of successes occur. Interested in probability of the k^{th} success occuring on the x^{th} trial, where x is the random variable. — x trials needed till k^{th} success
- **Geometric** $\sim Geometric(p)$**:** Negative binomial distribution with $k = 1$, stop after 1^{st} success. — x trials needed till 1^{st} success
- **Poisson** $\sim P(\lambda)$**:** Yield the number of successes occuring during a given time interval. *Poisson process* with rate parameter α are:
  - λ is the number of expected outcomes
  - expected number of occurences in interval of length T is αT
  - there are no simultaneous occurences
  - no. of occurrences in disjoint time intervals are independent
  - Approx. to $B(n, p)$: $\lim_{\substack{p \to 0 \\ n \to \infty}} \Pr(X = x) = \frac{e^{-np}(np)^x}{x!}$

Continuous Distributions — Formula Sheet

- **Uniform** $\sim U(a, b)$**:** Has uniform distribution over interval $[a, b]$ where $-\infty < a < b < \infty$, probability density function given by $f_X(x) = \frac{1}{b-a}, axb$, and 0 otherwise.
- **Exponential** $\sim Exp(\lambda)$**: non-negative** values is said to have exponential distribution with parameter $\lambda > 0$.
  - **No Memory:** $\Pr(X > s + t \mid X > s) = \Pr(X > t)$.
- **Normal** $\sim N(\mu, \sigma^2)$**:** symmetric about $x = \mu$, maximum point is at μ and its value is $\frac{1}{\sqrt{2\pi}}$. Total area under curve is 1, as σ increases, curve flattens. Standardized normal $= Z = \frac{(X-\mu)}{\sigma}$
  - Some statistical tables give the 100α percentage points, z_α, of a standardized normal distribution, where $\alpha = \Pr(Zz_\alpha) = \int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})dx$. $\rightarrow$ $\Pr(Zz_\alpha) = \Pr(Z - z_\alpha) = \alpha$
  - $\Phi(z) = \Pr(Zz)$, $1 - \Phi(z) = \Pr(Z > z)$
  - Approx. to $B(n, p)$ as $n \to \infty$ and $p \to \frac{1}{2}$, generally can approximate when $np > 5$ and $n(1 - p) > 5$: $Z = \frac{X-np}{\sqrt{npq}}$
  - **Continuity Correction:** When approximating binomial using normal, $Pr(X = k) \approx Pr(k - \frac{1}{2} < X < k + \frac{1}{2})$

Chapter 5

Population: Totality of all possible outcomes/observations of survey or experiment. **Sample:** Subset of a population **SRS:** every subset of n observations of the population has *same probability* of being selected. **Statistic:** A function of a random sample $(X_1, X_2, \ldots, X_n)$ is called a statistic e.g. *mean*.

Sample Mean: For some random sample of size n represented by $X_1, X_2, \cdots, X_n$, the sample mean is defined by the statistic $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. If the values in the random sample are observed and they are $x_1, x_2, \cdots, x_n$, then the realization of the statistic $\overline{X}$ is given by $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
Sample Variance: The sample variance, defined as $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is a statistic, if the values are observed to be $x_1 \ldots x_n$, realization of the statistic S^2 is given by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
Sampling Distribution: For random samples of size n taken from an infinite population or a finite population with replacement having population mean μ and population standard deviation σ, the sampling distribution of the sample mean $\overline{X}$ has its mean and variance given by:
1. $\mu_{\overline{X}} = \mu_X$, i.e. $E(\overline{X}) = E(X)$. 2. $\sigma_{\overline{X}}^2 = \frac{\sigma_X^2}{n}$, i.e. $V(\overline{X}) = \frac{V(X)}{n}$.

Law of Large Numbers: $X_1, X_2, \cdots, X_n$ be a random sample of size n from a population having any distribution with mean μ and **finite** population variance σ^2. For any $\epsilon \in \mathbb{R}$, $\text{P}(|\overline{X} - \mu| > \epsilon) \to 0$ as $n \to \infty$.
Central Limit Theorem: The sampling distribution of the sample mean $\overline{X}$ is approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ if n is sufficiently large. Hence $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ follows approximately $N(0, 1)$.
1. Central Tendency: $\mu_{\overline{X}} = \mu$. 2. Variation: $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

Chi-square Distribution $\sim \chi^2(n)$ Let $Z_1, \ldots, Z_n$ be n independent and identically distributed *standard normal* random variables. A random variable with the same distribution as $Z_1^2 + \ldots + Z_n^2$ is called a χ^2 random variable with n degrees of freedom.
1. For $Y \sim \chi^2(n)$ then $E(Y) = n$ and $V(Y) = 2n$

2. For large n, $\chi^2(n)$ is approximately $N(n, 2n)$
3. Y_1, Y_2 are *independent* χ^2 random variables with m, n *df*, then $Y_1 + Y_2$ is a χ^2 random variable with $m + n$ *df*.
4. All density functions have long right tail.
5. Define $\chi^2(n; \alpha)$ such that $Y \sim \chi^2(n)$ s.t $Pr(Y > \chi^2(n; \alpha)) = \alpha$
6. **Sampling Distribution:** $\frac{(n-1)S^2}{\sigma^2}$. If S^2 is variance of random sample of size n taken from a normal population with σ^2 then, random variable $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$ has χ^2 distribution with $n - 1$ df.

Degrees of freedom (*df*)	Significance level (α) .99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

***t*-distribution:** Suppose $Z \sim N(0, 1)$ and $U \sim \chi^2(n)$. If Z and U are independent then $T = \frac{Z}{\sqrt{U/n}}$ follows *t-distribution* with n df.

1. *t-distribution* with n df denoted by $t(n)$, approaches $N(0, 1)$ as parameter $n \to \infty$. When $n \geq 30$ we can replace it with $N(0, 1)$.
2. If $T \sim t(n)$, then $E(T) = 0$ and $V(T) = \frac{n}{n-2}$ for $n > 2$
3. Graph of $t(n)$ resembles graph of $N(0, 1)$, $Pr(T > t_{n;\alpha}) = \alpha$
4. If $X_1, \ldots X_n$ are independent and identically distributed normal random variables with μ, σ^2 then $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ follows a t-dist with $n - 1$ df.

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

F-distribution $\sim F(n_1, n_2)$: Suppose $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$ are *independent*, distirbution of random variable $F = \frac{U/m}{V/n}$ is an F-distribution with (m, n) df.

1. Denoted by $F(m, n)$, if $X \sim F(m, n)$ then $E(X) = \frac{n}{n-2}$ for $n > 2$ and $V(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ for $n > 4$
2. If $F \sim F(m, n)$ then $1/F \sim F(m, n)$, follows from def$^{\underline{n}}$ of F-distr.
3. Values in F: $F(m, n; \alpha)$ s.t. $P(F > F(m, n; \alpha)) = \alpha$, $F \sim F(m, n)$
4. It can be sown that $F(m, n; 1 - \alpha) = 1/F(n, m; \alpha)$

Chapter 6

Statistic: Function of random sample, not dependent on unknown parameters.

Estimator: Rule, usually expressed as a formula, indicate how to calculate an estimate based on information in the sample. For example, $\overline{X}$ is an estimator of μ. The value of $\overline{X}$, denoted by $\overline{x}$, is an estimate of μ. **Unbiased Estimator:** A statistic $\widehat{\Theta}$ is said to be an unbiased estimator of the parameter θ if $E(\widehat{\Theta}) = \theta$.

Maximum Error of Estimate: Typically $\overline{X} \neq \mu$, so $\overline{X} - \mu$ measures the difference between the estimator and the true value of the parameter. If the population is normal or if n is large, $\frac{\overline{X}-\mu}{\sigma\sqrt{n}}$ follows an approximately standard normal distribution.

$\to Pr(\left|\overline{X} - \mu\right| \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$, there is a probability error $\left|\overline{X} - \mu\right| \leq E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, aka **maximum error of estimate**.

Confidence Interval: An **interval estimator** is a rule for calculating, from the sample, an interval (a, b) in which you are fairly certain the parameter of interest lies in: $Pr(a < \mu < b) = 1 - \alpha$.

Pooled Estimator:

σ^2 can be estimated by: $S_p = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ with S_i^2 being the sample variances of either samples. This follows a t-distribution with degrees of freedom $n_1 + n_2 - 2$: $T = \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$

Paired Data:

1. $(X_1, Y_1), \cdots, (X_n, Y_n)$ are matched pairs, where $X_1, \cdots, X_n$ is a random sample from population 1, $Y_1, \cdots, Y_n$ is a random sample from population 2.
2. X_i and Y_i are *dependent.*
3. (X_i, Y_i) and (X_j, Y_j) are independent for any $i \neq j$.
4. For matched pairs, define $D_i = X_iY_i, \mu_D = \mu_1\mu_2$.
5. Now we can treat $D_1, D_2, \cdots, D_n$ as a random sample from a single population with mean μ_D and variance σ_D^2.

Point estimate of μ_D is $\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)$ Point estimate of variance σ_D^2 is given by $s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2$

$$\boxed{\Pr(-t_{n-1;\alpha/2} < T < t_{n-1;\alpha-2}) = 1 - \alpha}$$

where $T = \frac{\overline{d}-\mu_D}{s_d/\sqrt{n}}$ t_{n-1} distribution. We have a $(1 - \alpha)100\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$

$$\boxed{\overline{d} - t_{n-1;\alpha/2}(\frac{S_D}{\sqrt{n}}) < \mu_D < \overline{d} + t_{n-1;\alpha/2}(\frac{S_D}{\sqrt{n}})}$$

Sufficiently large sample $n > 30$, can replace $t_{n-1;\alpha/2}$ by $z_{\alpha/2}$ and get

$$\boxed{\overline{d} - z_{\alpha/2}(\frac{S_D}{\sqrt{n}}) < \mu_D < \overline{d} + z_{\alpha/2}(\frac{S_D}{\sqrt{n}})}$$

Chapter 7

Null and Alternative Hypotheses

Null hypothesis H_0: Formulated hypothesis with the hope of rejecting. Rejection of $H_0 \to$ acceptance of an alternative hypothesis, denoted by H_1. Reject hypothesis: conclude it is false. Accept hypothesis: insufficient evidence to believe otherwise.

	Do not reject H_0	Reject H_0
H_0 True	Correct Decision	**Type I Error**
H_0 False	**Type II Error**	Correct Decision

Significance Level vs Power

Level of Significance α: Probability of making a Type I error, $P(Type\ I\ Error) = P(Reject\ H_0|H_0\ true)$ **Power of the test** β: Define $1 - \beta = P(Reject\ H_0|H_0\ false)$ to be power of test.

Hypothesis testing concerning mean

- **Known Variance (Two-sided) - Critical Value:** 1. Variance, σ^2, known, AND 2. Underlying distribution is normal OR $n > 30$. Test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.
  1. $\overline{x}_1 < \overline{X} < \overline{x}_2$ OR $-z_{\alpha/2} < z < z_{\alpha/2}$defines the acceptance region
  2. 2 tails, $\overline{X} < \overline{x}_1$, $\overline{X} > \overline{x}_2$ constitute the critical/rejection region. We need $\overline{x}_1 = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $\overline{x}_2 = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. If $\overline{X}$ falls in the acceptance region, we *accept the null hypothesis*, else reject. The critical region is often stated in terms of Z instead of $\overline{X}$, where $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ $N(0, 1)$
- **Known Variance (Two-sided) - p-Value:** Probability of obtaining test statistic more extreme than the observed sample value given H_0 is true. aka observed level of significance.
  1. Sample statistic $(\overline{X})$ into test statistic (Z), obtain p-value
  2. Compare p with $\alpha/2$. If $p < \alpha/2$, reject H_0, else do not reject.
- **Known Variance (One-sided) - Critical Value:** Same as before but alternative hypothesis is now either $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$. In both cases, let $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$. Need to check if the observed values of Z is $> z_\alpha$ or $< -z_\alpha$ respectively.
- **Known Variance (One-sided) - p-Value:** Same as the two-sided known variance approach, just that we will compare against the relevant side, and against α itself.
- **Unknown Variance (Two-sided) - Critical Value:** 1. Variance unknown, AND 2. Underlying distribution is *normal.* Let $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$, where S^2 is the sample variance. Then H_0 is rejected if the observed value of T, say t, $> t_{n-1;\alpha/2}$ or $< -t_{n-1;\alpha/2}$.
- **Unknown Variance (One-sided) - Critical Value:** We test the relevant side, $t > t_{n-1;\alpha}$ or $t < -t_{n-1;\alpha}$. When $n \geq 30$ can replace t_{n-1} by $Z \sim N(0, 1)$.

Hypotheses Testing Concerning Difference Between Two Means

- **Known Variances:** 1. ariances σ_1^2 and σ_2^2 are known and 2. Underlying distribution is normal or both $n_1 30, n_2 30$. Generally, since variance is known, we will be using the Z distribution. For the null hypothesis H_0: $\mu_1 - \mu_2 = \delta_0$, test statistic given by: $Z = \frac{(\overline{X}-\overline{Y})-\delta_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$.
- **Unknown Variances (Large Samples):** 1. Variances σ_1^2 and σ_2^2 are unknown and 2. Both $n_1 30, n_2 30$. For the null hypothesis H_0: $\mu_1 - \mu_2 = \delta_0$, test statistic given by: $Z = \frac{(\overline{X}-\overline{Y})-\delta_0}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim N(0, 1)$.

H_1	Rejection Region	p-value
$\mu_1 - \mu_2 > \delta_0$	$z > z_\alpha$	$P(Z > \|z\|)$
$\mu_1 - \mu_2 < \delta_0$	$z < -z_\alpha$	$P(Z < -\|z\|)$
$\mu_1 - \mu_2 \neq \delta_0$	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	$2P(Z > \|z\|)$

- **Unknown but Equal Variances (Small Samples):** 1. Variances σ_1^2 and σ_2^2 are unknown but equal and 2. The populations are normal and 3. Both are small samples $n_1 30, n_2 30$. For the null hypothesis H_0: $\mu_1 - \mu_2 = \delta_0$, test statistic given by: $Z = \frac{(\overline{X}-\overline{Y})-\delta_0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$.

Paired Data For paired data, define $D_i = X_i - Y_i$, null hypothesis H_0: $\mu_D = \mu_{D0}$, test statistic given by: $T = \frac{\overline{D}-\mu_{D_0}}{S_D/\sqrt{n}}$.

- $n < 30$ and population normally distributed, then $T \sim t_{n-1}$
- $n \geq 30$ then $T \sim N(0, 1)$

$$\boxed{\text{We can roughly assume equal variance if } \tfrac{1}{2} \leq \tfrac{S_1}{S_2} \leq 2 \text{ as statistic not overly sensitive to small differences between population variances.}}$$

By parts: $\int uvdx = u\int vdx - \int \frac{du}{dx}(\int vdx)dx$

Subst: $I = \int f(x)dx$, $x = g(t) \to dx = g'(t)dt \mid I = \int f(g(t))g'(t)dt$

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample with population mean μ and variance $\sigma^2 > 0$. Denote by S^2 the sample variance based on this random sample, and denote by s^2 the computed value of the sample variance based on one set of observed values of $\{X_1, X_2, \ldots, X_n\}$. Which of the following statement is **IMPOSSIBLE**?

(a) $E(S^2) = \sigma^2$.
(b) $E(S) > \sigma$.
(c) $s^2 > \sigma^2$.
(d) $s^2 < \sigma^2$.

SOLUTION
The answer is (b).
Note that $[E(S)]^2 \leq E(S^2) = \sigma^2$. So (b) cannot be true.
(a) is always true (stated in lectures).
(c) and (d) is possible since they are observed sample variances.