STATS200 class notes

Erich Trieschman

2021 Fall quarter

Contents

1 Review: Combinatorics and probability **2**
1.1 Calculus cheat sheet . . . 2
1.2 Events and sets . . . 3
1.3 Probability . . . 3
1.3.1 Conditional probability . . . 3
1.3.2 Independence . . . 3

2 Random variables and common distribution functions **4**
2.1 Discrete distribution functions . . . 4
2.1.1 Bernoulli . . . 4
2.1.2 Binomial distribution . . . 4
2.1.3 Geometric distribution . . . 4
2.1.4 Negative binomial . . . 4
2.1.5 Poisson distribution . . . 4
2.2 Continuous distribution functions . . . 5
2.2.1 Uniform distribution . . . 5
2.2.2 Normal distribution . . . 5
2.2.3 Exponential distribution . . . 5
2.2.4 Gamma distribution . . . 5
2.2.5 Cauchy distribution . . . 5
2.2.6 Beta distribution . . . 6

3 Joint, marginal, and conditional distributions **6**
3.1 Joint distributions . . . 6
3.1.1 Distribution of $X + Y$ . . . 6
3.1.2 Expectation of joint distributions . . . 6
3.2 Marginal distributions . . . 7
3.3 Conditional distributions . . . 7
3.3.1 Transformations of random variables . . . 7

4 Expected variables **7**
4.1 Expected value . . . 7
4.2 Variance . . . 8
4.3 Covariance . . . 8
4.4 Correlation . . . 8
4.5 Key theorems . . . 9
4.5.1 Iterated expectation . . . 9
4.5.2 Variance decomposition . . . 9
4.5.3 Cauchy-Schwartz inequality . . . 9
4.5.4 Jensen inequality . . . 9
4.5.5 Markov inequality . . . 9
4.5.6 Chebyshev inequality . . . 9
4.6 Moment generating function . . . 9
4.6.1 Common MGF derivations . . . 10

5 Convergence and limit theorems **10**
5.1 Convergence in probability . . . 10
5.2 Convergence in L_p . . . 10
5.3 Convergence in distribution . . . 10
5.3.1 Convergence in probability $\implies$ convergence in distribution . . . 10
5.3.2 Slutsky's theorem . . . 11
5.3.3 Student's t distribution (example use case of Slutsky) . . . 11
5.4 Law of large numbers . . . 11
5.5 Central limit theorem . . . 11
5.5.1 Useful CLT properties . . . 12
5.6 Delta method . . . 12

6 Estimation **12**
6.1 Mean Squared Error . . . 12
6.2 Method of Moments estimator . . . 12
6.3 Maximum likelihood estimator . . . 13
6.4 Fisher Information . . . 13
6.4.1 Properties of Fischer Information . . . 13
6.4.2 The "Big" theorem: Asymptotic distribution using Fischer Information . . . 14
6.5 Bayes estimator . . . 14
6.5.1 Example Bayes estimator method . . . 15
6.5.2 Bayes estimator properties . . . 15
6.6 Sufficiency . . . 15
6.6.1 Fischer's Factorization Theorem . . . 15
6.6.2 Rao-Blackwell Theorem . . . 15

7 Hypothesis testing **15**
7.1 Likelihood ratio . . . 16
7.2 Neyman-Pearson lemma . . . 16
7.3 Uniformly Most powerful test (UMP) . . . 16
7.3.1 Example LR and UMP test . . . 17
7.4 P-values . . . 17
7.5 Generalized Likelihood Ratio test . . . 17
7.5.1 GLR example I . . . 17
7.5.2 GLR example II . . . 17
7.5.3 Testing multinomial distributions . . . 18

8 Helpful applied methods **18**
8.1 Derivation of Linear regression . . . 18
8.2 Confidence intervals . . . 18
8.3 Comparing two samples . . . 18
8.4 Fischer's Exact test . . . 18

1 Review: Combinatorics and probability

1.1 Calculus cheat sheet

Logs: $log_b(M*N) = log_bM + log_bN \bullet log_b(\frac{M}{N}) = log_bM - log_bN \bullet log_b(M^k) = klog_bM \bullet e^n e^m = e^{n+m}$

Derivatives: $(x^n)' = nx^{n-1} \bullet (e^x)' = e^x \bullet (e^{u(x)})' = u'(x)e^x \bullet (log_e(x))' = (lnx)' = \frac{1}{x} \bullet (f(g(x)))' = f'(g(x))g'(x)$

Integrals: $\int_a^b f(x)dx = \int_{g(a)}^{g(b)} f(g(u))g'(u)du$ where $g(u) = x \bullet \int_a^b u(x)v'(x)dx = u(b)v(b) - u(a)v(a) - \int_a^b u'(x)v(x)dx$

Infinite series and sums: $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!} \bullet (1 + \frac{a}{n})^n \longrightarrow e^a$
$ln(1+x) = 1 - x + \frac{x^2}{2} - \frac{x^3}{3} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^n}{n} \bullet \frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots = \sum_{n=0}^{\infty} a^x$ for $|x| < 1$

1.2 Events and sets

Set operations follow commutative, associative, and distributive laws:

- Commutative: $E \cup F = F \cup E$ and $E \cap F = F \cap E$ (also written $EF = FE$)
- Associative: $(E \cup F) \cup G = E \cup (f \cup G)$ and $(E \cap F) \cap G = E \cap (F \cap G)$
- Distributive: $(E \cup F) \cap G = (E \cap G) \cup (F \cap G) = E \cap G \cup F \cap G$ and $E \cap F \cup G = (E \cup G) \cap (F \cup G) = E \cup G \cap F \cup G$

DeMorgan's Laws relate the complement of a union to the intersection of complements:

$$(\cup_{i=1}^{n} E_i)^c = \cap_{i=1}^{n} E_i^c \bullet (\cap_{i=1}^{n} E_i)^c = \cup_{i=1}^{n} E_i^c$$

1.3 Probability

A **probability space** is defined by a triple of objects $(S, \mathcal{E}, P)$:

- S : Sample space
- $\mathcal{E}$: Set of possible events within the sample space. Set of events are assumed to be θ-field (below)
- P : Probability for each event

A **θ-field** is a collection of subsets $\mathcal{E} \subset S$ that satisfy $0 \in \mathcal{E} \bullet E \in \mathcal{E} \Rightarrow E^C \in \mathcal{E} \bullet E_i \in \mathcal{E}$ for $1, 2, \cdots \Rightarrow \cup_{i=1}^{\infty} E_i \in \mathcal{E}$

Probability properties:
$P(A^C) = 1 - P(A)$ $\bullet P(0) = 0$ $\bullet A \subset B \longrightarrow P(A) \leq P(B)$ $\bullet P(A \cup B) = P(A) + P(B) - P(A \cap B)$

The **law of total probability** relates marginal probabilities to conditional probabilities. For a partition, $E_1, E_2, \ldots$ of set, S, where a partition implies i) E_i, E_j are pairwise disjoint and ii) $\cup_{i=1}^{\infty} E_i = S$, then

$$P(A) = \sum_{i=1}^{\infty} P(A \cap E_i) = \sum_{i=1}^{\infty} P(A \mid E_i) P(E_i)$$

The **continuity of probability measures** state

$$(i)\ E_1 \subset E_2 \subset \ldots \text{ Let } E_\infty = \cup_i E_i, \text{ then } P(E_n) \longrightarrow P(E_\infty) \text{ as } n \longrightarrow \infty$$
$$(ii)\ E_1 \supset E_2 \supset \ldots \text{ Let } E_\infty = \cap_i E_i, \text{ then } P(E_n) \longrightarrow P(E_\infty) \text{ as } n \longrightarrow \infty$$

1.3.1 Conditional probability

The conditional probability is the probability of one event occurring, given the other event occurring. A reframing of conditional probability (see formula below) is the probability of both events occurring, divided by the marginal probability of one of the events occurring.

$$p_{X|Y}(x|y) = \frac{p_{x,y}(x, y)}{p_y(y)}$$

Bayes Theorem leverages conditional probabilities of measured events to glean conditional probabilities of unmeasured events:

$$P(E_i \mid B) = \frac{P(B \mid E_i) P(E_i)}{\sum_{j=1}^{\infty} P(B \mid E_j) P(E_j)} = \frac{P(B \mid E_i) P(E_i)}{P(B)}$$

Where $E_1, E_2, \ldots$ form a partition of the sample space.

1.3.2 Independence

Events A and B are independent if $P(A \cap B) = P(A)P(B)$

It is possible for events to be pairwise independent, but not mutually independent. For example, toss a pair of dice and let D_1 be the number for die 1 and D_2 be the number for die 2. Define $E_i = \{D_i \leq 2\}$. And define $E_3 = \{3 \leq \max(D_1, D_2) \leq 4\}$. These events are pairwise independent, but $P(E_1 \cap E_2 \cap E_3) = 0$, so they are not mutually independent.

2 Random variables and common distribution functions

Random variables are functions connecting a sample space to real numbers: $\{\omega \in S : X(\omega) \leq t\} \in \mathcal{E}$. For example, if coin tosses produce a sample space of {Heads, Tails}, a random variable can be the number of heads.

2.1 Discrete distribution functions

2.1.1 Bernoulli

Probability mass function ($Bernouli(p)$): Random variable X takes the value 1 with probability p and the value 0 with probability $1-p$

$$p(x) = p^x(1-p)^{1-x} \ , \ x \in \{0,1\}$$

Expected value: p
Variance: $p(1-p)$

2.1.2 Binomial distribution

Probability mass function ($Bin(n,p)$): For random variable X, the number of successes in n trials, the probability of observing j successes where each success has probability p is

$$P(X=j) = \binom{n}{j} p^j (1-p)^{n-j}$$

Expected value: np
Variance: $np(1-p)$
MLE: $\hat{p} = X/n$

2.1.3 Geometric distribution

Probability mass function ($Geom(p)$): For random variable X, the number of trials until the first success (included) with probability p is

$$P(X=j) = (1-p)^{j-1}p$$

Expected value: $\frac{1}{p}$
Variance: $\frac{1-p}{p}$

2.1.4 Negative binomial

Probability mass function ($NB(r,p)$): For random variable X, the number of successes, k before a specified number of failures, r, with probability of success p is

$$P(X=k) = \binom{k+r-1}{k} (1-p)^r p^k$$

Expected value: $\frac{pr}{1-p}$
Variance: $\frac{pr}{(1-p)^2}$

2.1.5 Poisson distribution

Probability mass function ($Pois(\lambda)$): For random variable, X, the number of events, k, occurring in a fixed interval of time or space if these events occur with a known constant mean rate, λ

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Expected value: λ
Variance: λ
MLE: $\hat{\lambda} = \bar{X}$

2.2 Continuous distribution functions

2.2.1 Uniform distribution

$Unif(a, b)$**:** The distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds.[1] The bounds are defined by the parameters, a and b, which are the minimum and maximum values

$$pdf: \ f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases} \bullet \ cdf: \ F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b] \\ 1 & \text{for } x > b \end{cases};$$

Expected value: $\frac{1}{2}(a+b)$
Variance: $\frac{1}{12}(b-a)^2$
MLE: $\hat{\theta} = X_{(n)} = max\{X_1, \ldots, X_n\}$

2.2.2 Normal distribution

$N(\mu, \sigma)$

$$pdf: \ f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \bullet \ cdf: \ F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

Expected value: μ
Variance: σ^2
MLE: $\hat{\mu} = \hat{X}$ • $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

2.2.3 Exponential distribution

$Exp(\lambda)$**:** the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate. It is a particular case of the gamma distribution.

$$pdf: \ f(x) = \lambda e^{-\lambda x} \bullet \ cdf: \ F(x) = 1 - e^{-\lambda x}$$

Expected value: $\frac{1}{\lambda}$
Variance: $\frac{1}{\lambda^2}$
MLE: $\hat{\lambda} = 1/\bar{X}$

2.2.4 Gamma distribution

$Gamma(\alpha, \lambda)$**:** a two-parameter family of continuous probability distributions.

$$pdf: f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \text{ , where } \Gamma(\alpha) = (\alpha-1)! \text{ for any positive integer, } \alpha$$

$$cdf: F(x) = \frac{1}{\Gamma(\alpha)} \gamma(\alpha, \lambda x), \text{ where } \gamma(\alpha, x) = \int_0^x t^{\alpha-1} e^{-t} dt$$

Expected value: $\frac{\alpha}{\lambda}$
Variance: $\frac{\alpha}{\lambda^2}$

2.2.5 Cauchy distribution

$Cauchy(t, s)$**:** The Cauchy distribution is often used in statistics as the canonical example of a "pathological" distribution since both its expected value and its variance are undefined

$$pdf: f(x) = \frac{1}{s\pi(1 + (x-t)/s)^2)}, \text{ where } s \text{ is the scale parameter and } t \text{ is the location parameter}$$

$$cdf: \frac{1}{\pi} \arctan\left(\frac{x-t}{s}\right) + \frac{1}{2}$$

Expected value: DNE
Variance: DNE

2.2.6 Beta distribution

Beta(α, β)**:** a family of continuous probability distributions defined on the interval [0, 1] parameterized by two positive shape parameters that appear as exponents of the random variable and control the shape of the distribution.

$$pdf : f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \text{ where } x \in [0,1], \text{ and } \Gamma(k) = (k-1)! \text{ for any positive integer } k$$

Expected value: $\frac{\alpha}{\alpha+\beta}$
Variance: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

3 Joint, marginal, and conditional distributions

3.1 Joint distributions

General case:

$$\text{cdf: } F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P(X_i \le x_1,\ldots,X_n \le x_n) \iff P((X_1,\ldots,X_n) \in E) = \int\cdots\int_E f_{X_1,\ldots,X_n} dx_1 \ldots dx_n$$

$$\text{pmf: } f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P(X_1 = x_1,\ldots,X_n = x_n)$$

When X_i independent:

$$\text{cdf: } P(X_1 \le x_1,\ldots,X_n \le x_n) = P(X_1 \le x_1)\ldots P(X_n \le x_n) = \prod_{i=1}^{n} P(X_i \le x_i)$$

$$\text{pmf: } P(X_1 = x_1,\ldots,X_n = x_n) = P(X_1 = x_1)\ldots P(X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)$$

3.1.1 Distribution of $X + Y$

The distribution of a sum of random variables is called a **convolution**. For X, Y independent

$$\begin{aligned}
F_{X+Y}(t) &= P(X+Y \le t) = P(X \le t-y) \\
&= \int_{-\infty}^{\infty} P(X \le t-y \mid Y = y) f_x(y) dy, \text{ to get marginal distribution} \\
&= \int_{-\infty}^{\infty} F_x(t-y) f_Y(y) dy, \text{ since } X, Y \text{ independent} \\
f_{X+Y}(t) &= \int_{-\infty}^{\infty} f_x(t-y) f_x(y) dy \\
p_{X+Y}(t) &= P(X+Y=t) = \sum_{x=-\infty}^{\infty} p_X(t-y) p_Y(y)
\end{aligned}$$

3.1.2 Expectation of joint distributions

For X, Y joint distribution, $f_{X,Y}(x,y)$, or probability mass function, $p(x,y)$

$$\text{pmf: } E[g(X,Y)] = \sum_s g(X(s),Y(s))p(s) = \sum_x \sum_y g(x,y) \sum_{s:X(s)=x,Y(s)=y} p(s) = \sum_x \sum_y g(x,y)p(x,y)$$

$$\text{pdf: } E[g(X,Y)] = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} g(x,y) f(x,y) dx dy$$

3.2 Marginal distributions

Marginal density functions or marginal probability mass functions are obtained by integrating or summing out the other variables

$$pmf : p_Y(y) = \sum_x yP(Y = y \mid x) \bullet pdf : F_Y(y) = \int_a^b f(x, y)dx, \text{ where } x \in [a, b]$$

3.3 Conditional distributions

Law of total probability:

$$P(E) = \sum_{i=-\infty}^{\infty} P(E \mid X = x)P(X) \text{ and } P(E) = \int_{-\infty}^{\infty} P(E \mid X = x)f(x)dx$$

$$\text{Recall: } p_{X|Y}(x|y) = \frac{p_{x,y}(x, y)}{p_y(y)} \text{ and } f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

3.3.1 Transformations of random variables

For X with density f_X and $Y = g(X)$

$$F_Y(y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$
$$f_Y(y) = \frac{d}{dy}F_X(g^{-1}(y)) = f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|$$

4 Expected variables

4.1 Expected value

$$E(X) = \sum_x xP(X = x)$$

Which can also be written as

$$E(X) = \sum_{x \in S} X(s)p(s), \text{ where } p(s) \text{ is the probability that element } s \in S \text{ occurs:}$$

Proof:

$$E(X) = \sum_i x_i P(X = x_i), \text{ for } E_i = \{X = x_i\} = \{s \in S : X(s) = x_i\}$$
$$= \sum_i x_i \sum_{s \in E_i} p(s) = \sum_i \sum_{s \in E_i} x_i p(s) = \sum_i \sum_{s \in E_i} X(s)p(s) = \sum_{s \in S} x_i p(s)$$

This equation structure helps proof several properties of the expected value:

- $E(g(X)) = \sum_i g(x_i)p_X(x_i)$, assuming $g(x_i) = y_i$

$$\sum_i g(x_i)p_X(x_i) = \sum_j \sum_{i:g(x_i)=y_j} g(x_i)p_X(x_i) = \sum_j \sum_{i:g(x_i)=y_j} y_j p_X(x_i) = \sum_j y_j P(g(X) = x_i) = E(g(X))$$

- $E(aX + b) = aE(X) + b$

$$E(aX + b) = \sum_{s \in S}(aX(s) + b)p(s) = a\sum_{s \in S} X(s)p(s) + \sum_{s \in S} bp(s) = aE(X) + b$$

- $E(X + Y) = E(X) + E(Y)$

$$E(X + Y) = \sum_{s \in S}(X(s) + Y(s))p(s) = \sum_{s \in S} X(s)p(s) + \sum_{s \in S} Y(s)p(s) = E(X) + E(Y)$$

4.2 Variance

$$Var(X) = E((X - E(X)))^2) = \sigma^2 \bullet SD = \sqrt{Var(X)} = \sqrt{\sigma^2} = \sigma$$

Several properties of variance follow from linearity of expectation:

(*i*) $Var(X) = E(X^2) - \mu^2$
$Var(X) = E((X-\mu)^2) = E(X^2 - 2X\mu + \mu^2) = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2$

(*ii*) $Var(aX + b) = a^2 Var(X)$
$Var(aX + b) = E((aX + b)^2) - E(aX + b)^2 = E(a^2X^2 + 2abX + b^2) - (aE(X) + b)^2$
$Var(aX + b) = a^2E(X^2) + 2abE(X) + b^2 - a^2E(X)^2 - 2abE(X) - b^2 = a^2E(X^2) - a^2E(X)^2 = a^2(E(X^2) - E(X)^2)$

(*iii*) $Var(X + Y) = Var(X) + Var(Y)$ for X, Y independent
$Var(X + Y) = E((X + Y)^2) - E(X + Y)^2 = E(X^2) + 2E(XY) + E(Y^2) - E(X^2) - 2E(X)E(Y) - E(Y)^2$
$Var(X + Y) = E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2$, since $E(XY) = 0$ (by independence) and $E(X) = E(Y) = 0$ (WLOG)
$Var(X + Y) = Var(X) + Var(Y)$

4.3 Covariance

$$Cov(X, Y) = E((X - E(X)(Y - E(Y)) = E(XY) - E(X)E(Y)$$

Several properties of covariance follow from linearity of expectation

(*i*) $Cov(X, X) = Var(X)$:
$Cov(X, X) = E[(X - E(X)(X - E(X))] = E[(X - E(X))^2] = Var(X)$

(*ii*) $Cov(X, Y) = E(XY) - E(X)E(Y)$:
$Cov(X, Y) = E[(X - E(X)(Y - E(Y))] = E(XY - E(Y)X - E(X)Y + E(X)E(Y))$
$Cov(X, Y) = E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) = E(XY) - E(X)E(Y)$

(*iii*) if X, Y independent, then$Cov(X, Y = 0)$

(*iv*) $Cov(aX, bY) = abCov(X, Y)$:
$Cov(aX, bY) = E(abXY) - E(aX)E(bY) = ab(E(XY) - E(X)E(Y)) = abCov(X, Y)$

(*v*) $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$:
$Cov(X, Y + Z) = E(X(Y + Z)) - E(X)E(Y + Z)$
$Cov(X, Y + Z) = E(XY) + E(XZ) - E(X)E(Y) - E(X)E(Z) = Cov(X, Y) + Cov(X, Z)$

(*vi*) $Cov(U, V) = \sum_i \sum_j b_i d_j Cov(X_i, Y_j)$, with $U = a + \sum_i b_i X_i$ and $V = c + \sum_j d_j Y_j$:

(*vii*) $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$:
$Var(X + Y) = Cov(X + Y, X + Y) = Cov(U, V)$, for $U = V = X + Y$
$Var(X + Y) = Cov(U, V) = Cov(X, X) + Cov(X, Y) + Cov(Y, Y) + Cov(Y, X)$, using vi
$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

4.4 Correlation

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

4.5 Key theorems

4.5.1 Iterated expectation

Law of iterated expectation: $E(E(Y \mid X)) = E(Y)$
Proof:

$$E(Y \mid X) = \sum_y y \frac{f_{X,Y}(X, y)}{f_X(X)}$$

$$E(E(Y \mid X)) = \sum_x \sum_y \left(y \frac{f_{X,Y}(x, y)}{f_X(x)} \right) f_X(x) = \sum_x \sum_y y f_{X,Y}(x, y) = \sum_y y f_Y(y) = E(Y)$$

4.5.2 Variance decomposition

Variance decomposition formula: $Var(Y) = E(Var(Y \mid X)) + Var(E(Y \mid X))$

4.5.3 Cauchy-Schwartz inequality

Cauchy-Schwartz inequality: $E(UV)^2 \leq E(U^2)E(V^2)$, with equality if $P(cU = U) = 1$ for some constant, c
Proof:

$$\text{let } h(t) = E((tU - V)^2) \geq 0,\ h(t) = t^2 E(U^2) - 2tE(UV) + E(V^2), \text{ a quadradic equation}$$
$$h(t) \geq 0 \Rightarrow \text{discriminant} \leq 0 \iff 4E(UV)^2 - 4E(U^2)E(V^2) \leq 0 \iff E(UV)^2 \leq E(U^2)E(V^2)$$

4.5.4 Jensen inequality

Jensen inequality: $E(g(x)) \geq g(E(x))$ for $g(x)$ convex
Proof: Let $E(X) = \mu$, and $L(X)$ a line s.t. $L(\mu) = g(E(x))$:

$$g(X) \geq L(X) \text{ for all } X \iff E(g(X)) \geq E(L(X)) = L(E(X)) = g(E(X))$$

4.5.5 Markov inequality

Markov inequality: For $X \geq 0$, $P(X \geq t) \leq \frac{E(X)}{t}\ \ \forall t > 0$
Proof:

$$\text{Let } y = \begin{cases} 1 & X \geq t \\ 0 & \text{otherwise} \end{cases}, \text{ Then } tY \leq X \text{ since } \begin{cases} X \geq t & t * 1 \leq X \\ X < t & t * 0 < X \end{cases}$$

$$tY \leq X \Longrightarrow E(tY) \leq E(X) \Longrightarrow tP(X \geq t) \leq E(X)) \Longrightarrow P(X \geq t) \leq \frac{E(X)}{t}$$

4.5.6 Chebyshev inequality

Chebyshev inequality: $P(|X - E(X)| \geq t) \leq \frac{Var(X)}{t^2}\ \ \forall t > 0$
Proof:

$$P(|X - E(X)| \geq t) = P((X - E(X))^2 \geq t^2) \leq \frac{E((X - E(X))^2)}{t^2}, \text{ by Markov inequality}$$
$$P((X - E(X))^2 \geq t^2) \leq \frac{Var(X)}{t^2}$$

4.6 Moment generating function

The moment generating function of a random variable X is defined as

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{n=0}^{\infty} \frac{\mathbb{E}[X^n]}{n!} t^n \leftarrow \text{power series}$$

Notice its called a moment generating function because each derivative of this function can generate a new moment of X at $t = 0$:

$$M_X^{(n)}(0) = \mathbb{E}[X^n]$$

4.6.1 Common MGF derivations

- $Y = a + bX \implies M_Y = e^{at} M_X(bt)$
- $Z = X + Y, X \perp Y \implies M_Z = M_Y M_X = E(e^t X)E(e^t Y)$

5 Convergence and limit theorems

5.1 Convergence in probability

A sequence of random variables, X_n, converges in probability, $X_n \xrightarrow{p} X$ when $P(|X_n - X| > \epsilon) \longrightarrow 0$ as $n \longrightarrow \infty$

Consistent estimator: $T_n = T_n(X_1, \ldots, X_n)$ converges in probability to $g(\theta)$, a function of the model parameter

Additional properties of convergence in probability

- if $X_n \xrightarrow{p} X$ and $a_n \xrightarrow{p} a$ then $a_n X_n \xrightarrow{p} aX$
- if $X_n \xrightarrow{p} X$ and $A_n \xrightarrow{p} A$ then $A_n X_n \xrightarrow{p} AX$
- if $X_n \xrightarrow{p} X$, $A_n \xrightarrow{p} A$, and $B_n \xrightarrow{p} B$ then $A_n X_n + B_n \xrightarrow{p} AX + B$
- if $X_n \xrightarrow{p} X$ and g a continuous function then $g(X_n) \xrightarrow{p} g(X)$ **(continuous mapping theorem)**

5.2 Convergence in L_p

See https://en.wikipedia.org/wiki/Lp_space for more information (not much covered in class).
Convergence in L_p is stronger than convergence in probability. **Counter example** to convergence in probability $\implies$ convergence in L_p:

$$\text{Let } X_n = \begin{cases} n & \frac{1}{n} \\ 0 & 1 - \frac{1}{n} \end{cases}$$

$$X_n \xrightarrow{p} 0: \; P(|X_n - 0| \geq \epsilon) = P(X_n = n) = 1/n \longrightarrow 0 \text{ as } n \longrightarrow 0$$

$$\text{but } E(X_n) = n\frac{1}{n} + 0(1 - \frac{1}{n}) = 1 \implies \text{ no convergence in } L_p$$

5.3 Convergence in distribution

A sequence of random vectors, X_n, converges in distribution to a random vector, $X_n \xrightarrow{d} X$ when

$$\lim_{n \longrightarrow \infty} F_{X_n}(x) = F_X(x) \text{ at all continuity points in } F_X$$

- Convergence in distribution **does not** imply convergence in probability unless convergence in distribution is to a single point
- if $X_n \xrightarrow{d} X$ and g a continuous function then $g(X_n) \xrightarrow{d} g(X)$ **(continuous mapping theorem)**

5.3.1 Convergence in probability $\implies$ convergence in distribution

Let X have cdf, F, with t a continuity point of F

$$\begin{aligned} P(X_n \leq a) \leq & P(X \leq a + \epsilon) + P(|X_n - X| > \epsilon) \text{ by lemma} \\ P(X \leq a - \epsilon) - P(|X_n - X| > \epsilon) \leq & P(X \leq a) \leq P(X \leq a + \epsilon) + P(|X_n - X| > \epsilon) \\ F_X(a - \epsilon) \leq & \lim_{n \to \infty} P(X_n \leq a) \leq F_X(a + \epsilon), \text{ where } F_X(a) = P(X \leq a) \\ \implies & \lim_{n \to \infty} P(X \leq a) = P(X \leq a) \implies \{X_n\} \xrightarrow{d} X \end{aligned}$$

5.3.2 Slutsky's theorem

$A_nX_n + B_n \xrightarrow{d} aX + b$ if $\{X_n\}$ sequence with $X_n \xrightarrow{d} X$, $\{A_n\}$ sequence with $A_n \xrightarrow{d} A$, $\{B_n\}$ sequence with $B_n \xrightarrow{d} b$

5.3.3 Student's t distribution (example use case of Slutsky)

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\frac{\sigma}{\hat{\sigma}}, \text{ and we know } \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1) \text{ and } \frac{\sigma}{\hat{\sigma}} \xrightarrow{p} 1 \text{ since } \hat{\sigma} \xrightarrow{p} \sigma$$

$$\text{So, by Slutsky's theorem, } \frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}} \xrightarrow{d} N(0,1) * 1$$

This RHS term is referred to as the t-statistic, which follows a Student's t distribution with $n-1$ degrees of freedom. In practice, if the sample is reasonably sized, it won't make a difference using the Normal distribution instead of the Student's t distribution.

5.4 Law of large numbers

For $X_1, X_2, \ldots, X_n$ i.i.d. with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then for any $\epsilon > 0$

$$P(|\overline{X}_n - \mu| > \epsilon) \longrightarrow 0 \text{ as } n \to \infty$$

Proof:

$$\mathbb{E}(\overline{X}_n) = \frac{1}{n}\sum_{I=1}^n \mathbb{E}(X_i) = \mu$$

$$Var(\overline{X}_n) = \frac{1}{n^2}\sum_{I=1}^n Var(X_i) = \frac{\sigma^2}{n}, \text{ since } X_i \text{ independent}$$

$$P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{Var(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty, \text{ by Chebyshev inequality}$$

5.5 Central limit theorem

Most useful form of CLT, which can be used for approximate methods:

$$\sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma} \longrightarrow N(0,1) \Longleftrightarrow \sqrt{n}(\overline{X}_n - \mu) \longrightarrow N(0, \sigma^2)$$

Formal definition: For $X_1, X_2, \ldots, X_n$ i.i.d. with $E(X_i) = 0$ (WLOG), $Var(X_i) = \sigma^2$, c.d.f, F, and MGF, M, (defined in a neighborhood of zero). Then

$$\lim_{n\to\infty} P(\frac{S_n}{\sigma\sqrt{n}} \leq x) = \Phi(x), \text{ for } S_n = \sum_{i=1}^n X_i$$

Proof: Let $Z_n = \frac{S_n}{\sigma\sqrt{n}}$. We show the MGF of Z_n tends to the MGF of the standard normal distribution. Since S_n is a sum of independent random variables,

$$M_{S_n}(t) = [M(t)]^n \text{ and } M_{Z_n}(t) = [M(\frac{t}{\sigma\sqrt{n}})]^n$$

$$\text{Reminder: Taylor series expansion of } M(s) = M(0) + sM'(0) + \frac{1}{2}sM''(0) + \epsilon_s$$

$$M(\frac{t}{\sigma\sqrt{n}}) = 1 + \frac{1}{2}\sigma^2(\frac{t}{\sigma\sqrt{n}})^2 + \epsilon_n \text{ with } E(X) = M'(0) = 0, Var(X) = M''(0) = \sigma^2$$

$$M_{Z_n}(t) = (1 + \frac{t^2}{2n} + \epsilon_n)^n \longrightarrow e^{\frac{t^2}{2}} \text{ as } n \longrightarrow \infty, \text{ by the infinite series convergence to } e^a$$

Since $e^{\frac{t^2}{2}}$ is the MGF of the standard normal distribution, we have proven the central limit theorem.

5.5.1 Useful CLT properties

- $X_i \sim N(0,1) \Longrightarrow \sum_{i=1}^{n} X_i \sim N(0,n) \Longrightarrow \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(0, n/n^2) = N(0, 1/n)$

5.6 Delta method

If g is a differentiable function at μ, $\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} N(0, g'(\mu)^2\sigma^2)$

Proof: For general g and assuming $E(\bar{X}_n) = \mu$

$$g(\bar{X}_n) \approx g(\mu) + g'(\mu)(\bar{X}_n - \mu) + \frac{1}{2}g''(\mu)(\bar{X}_n - \mu)^2 + \epsilon \text{ (Taylor approximation of } g(\mu))$$

$$g(\bar{X}_n) - g(\mu) \approx g'(\mu)(\bar{X}_n - \mu) + \epsilon \Longleftrightarrow \sqrt{n}(g(\bar{X}_n) - g(\mu)) \approx g'(\mu)\sqrt{n}(\bar{X}_n - \mu) + \epsilon \text{ and we know}$$

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \Longleftrightarrow g'(\mu)\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, g'(\mu)^2\sigma(2))$$

So $\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} N(0, g'(\mu)^2\sigma(2))$

Note: if we find that $g'(\mu) = 0$, then repeat this process with the second derivative, $g''(\mu)$.

6 Estimation

Here we use functions of the data ("estimators"), $T(X_1, \ldots, X_n)$ to estimate population parameters, θ

6.1 Mean Squared Error

The **Mean Squared Error (MSE)** can be used to evaluate our estimators.

$$\begin{aligned} MSE(T,\theta) &= E_\theta[(T - g(\theta))^2] = E_\theta(T^2) - 2g(\theta)E_\theta(T) + g(\theta)^2 \\ &= Var_\theta(T) + E_\theta(T)^2 + 2g(\theta)E_\theta(T) + g(\theta)^2 = Var_\theta(T) + (E_\theta(T) - g(\theta))^2 \\ &= Var_\theta(T) + Bias_\theta^2(T), \text{ where } Bias_\theta(T) = E_\theta(T) - g(\theta) \end{aligned}$$

6.2 Method of Moments estimator

To generate a method of moments estimator

- Calculate a moment with MGF of the assumed distribution. Any moment, k, can be used, but lower moments will typically lead to an estimator distribution with lower variance

$$E(X^k) = g(\theta)$$

- Invert this expression to create an expression for the parameter(s) in terms of the moment

$$g^{-1}(E(X^k)) = \theta \Longrightarrow f(E(X^k)) = \theta, \text{ where } f(x) = g^{-1}(x)$$

- Insert the sample moment into this expression, thus obtaining estimates of the parameters in terms of data

$$\hat{\theta} = f(\frac{1}{n}\sum X_i^k) \text{ , by LNN } \frac{1}{n}\sum X_i^k \xrightarrow{p} E(X^k)$$

- Use the delta method to determine what the method of moments estimator converges to in distribution

$$\sqrt{n}(f(\frac{1}{n}\sum X_i^k) - \theta) \xrightarrow{d} N(0, f'(E(X_i^k))^2 Var(X_i^k)^2)$$

Methods of moment estimators are not uniquely determined, nor must they exist. The motivation for subsequent estimators is to help us pick the estimator with the smallest possible variance.

6.3 Maximum likelihood estimator

The **maximum likelihood estimator** constructs an estimator, $\hat{\theta}_{MLE}$, that maximizes the likelihood function with respect to θ.

The **likelihood function**, $L(\theta)$ is the joint density or probability mass function, $f(X, \theta)$, evaluated at the data, $\{X_i, \ldots, X_n\}$. Assuming the data is *i.i.d.*:

$$L(\theta) = \prod_{i=1}^{n} f(X_i, \theta)$$

General approach to constructing MLE:

- Construct the likelihood function: $L(\theta) = \prod_{i=1}^{n} f(X_i, \theta)$

  Example normal: $L(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right)$

  Example restricted multinomial: $L(\theta) \propto f_1(\theta)^{X_1} \ldots f_k(\theta)^{X_k}$
- Take the log of the likelihood: $log(L(\theta)) = l(\theta) = \sum_{i=1}^{n} log(f(X_i, \theta))$
- Take the derivative of the log-likelihood function with respect to θ: $\frac{d}{d\theta} l(\theta) = \sum_{i=1}^{n} \frac{d}{d\theta} log(f(X_i, \theta))$
- Find critical points of this function ($0 = \sum_{i=1}^{n} \frac{d}{d\theta} log(f(X_i, \hat{\theta}))$) and determine that one is a maximum ($0 = \sum_{i=1}^{n} \frac{d^2}{d\theta^2} log(f(X_i, \hat{\theta}))$, checking if $\hat{\theta} < 0$)

Approach to constructing MLE when Indicators are present:

- Simplify likelihood function (splitting indicators when possible)
- Make an argument for why the function is increasing or decreasing
- Determine the value at the bounds of the function

6.4 Fisher Information

The **information** that data, X, contains about parameter, θ is defined by

$$I(\theta) = E_\theta\left[\left(\frac{d}{d\theta} log(f(X, \theta))\right)^2\right]$$

Fisher Information assumes **differentability** and **existence of the second moment**. $\frac{d}{d\theta} log(f(X, \theta))$ is called the **score** function

6.4.1 Properties of Fischer Information

1. $E_\theta\left[\left(\frac{d}{d\theta} log(f(X, \theta))\right)\right] = 0$:

$$E_\theta\left[\left(\frac{d}{d\theta} log(f(X, \theta))\right)\right] = \int \frac{d}{d\theta} log(f(x, \theta)) f(x, \theta) dx = \int \frac{f'(x, \theta)}{f(x, \theta)} f(x, \theta) dx = \int f'(x, \theta) dx = \frac{d}{d\theta} \int f(x, \theta) dx = \frac{d}{d\theta} * 1 = 0$$

2. $I(\theta) = Var\left(\frac{d}{d\theta} log(f(X, \theta))\right)$:

$$Var\left(\frac{d}{d\theta} log(f(X, \theta))\right) = E_\theta\left[\left(\frac{d}{d\theta} log(f(X, \theta))\right)^2\right] - E_\theta\left[\left(\frac{d}{d\theta} log(f(X, \theta))\right)\right]^2 = I(\theta) - 0^2 = I(\theta)$$

3. $I(\theta) = -E_\theta\left[\frac{d^2}{d\theta^2} log(f(X, \theta))\right]$:

$$\frac{d}{d\theta} log(f(x, \theta)) = \frac{f'(x, \theta)}{f(x, \theta)} \implies \frac{d^2}{d\theta^2} log(f(x, \theta)) = \frac{f(x, \theta) f''(x, \theta) - f'(x, \theta)^2}{f(x, \theta)^2}$$

$$E\left[\frac{d^2}{d\theta^2} log(f(x, \theta))\right] = \int \frac{f(x, \theta) f''(x, \theta) - f'(x, \theta)^2}{f(x, \theta)^2} f(x, \theta) dx = \int f''(x, \theta) - I(\theta) = -I(\theta), \text{ since } \int \frac{d^2}{d\theta^2} f(x, \theta) = \frac{d^2}{d\theta^2} * 1 = 0$$

4. $I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta)$ for X, Y independent : (Information increases with larger sample!)

Corrolary: $I_n(\theta) = nI_1(\theta)$ for $X_1, \ldots, X_n$ *i.i.d* with $I_1(\theta)$ the Information based on one data

5. **Cramer-Rau-Fisher Inequality:** $Var(T(X)) \geq \frac{g'(\theta)^2}{I(\theta)}$ for $E(T(X)) = g(\theta)$:

$Cov[T(X), \frac{d}{d\theta} log(f(X,\theta))] = E[T(X)\frac{d}{d\theta} log(f(X,\theta))]$, using property 1

$Cov[T(X), \frac{d}{d\theta} log(f(X,\theta))] = \int T(x)f'(x,\theta)dx = \frac{d}{d\theta}\int T(x)f(x,\theta)dx = \frac{d}{d\theta}E(T(X)) = \frac{d}{d\theta}g(\theta) = g'(\theta)$

$g'(\theta)^2 \leq Var(T(X))Var\left(\frac{d}{d\theta} log(f(X,\theta))\right) = Var(T(X))I(\theta)$ by correlation inequality: $\rho^2 \leq 1$

$Var(T(X)) \geq \frac{g'(\theta)^2}{I(\theta)}$

6.4.2 The "Big" theorem: Asymptotic distribution using Fischer Information

Under regularity assumptions, the maximum likelihood estimator (or any other reasonable estimator), $\hat{\theta}$ of θ satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right)$$

Sketch of proof:

$$L(\theta) = \prod_{i=1}^{n} f(X_i, \theta) \Longleftrightarrow l(\theta) = log(L(\theta)) = \sum_{i=1}^{n} log(f(X_i, \theta))$$

MLE solves $l'(\hat{\theta})$ =0, with $l'(\theta) \approx\ l'(\theta_0) + (\hat{\theta} - \theta_0)l''(\theta_0)$ (full proof requires showing the error in this approx. is small)

$$0 = l'(\theta_0) + (\hat{\theta} - \theta_0)l''(\theta_0) \Longrightarrow \hat{\theta} - \theta_0 = \frac{l'(\theta_0)}{l''(\theta_0)} \Longleftrightarrow \sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}\frac{l'(\theta_0)}{l''(\theta_0)} = \frac{l'(\theta_0)}{\sqrt{n}} \div \frac{l''(\theta_0)}{n}$$

$$\frac{l''(\theta_0)}{n} = \frac{\sum \frac{d^2}{d\theta^2} log(f(X,\theta))}{n} \xrightarrow{p} -E_\theta\left[\frac{d^2}{d\theta^2} log(f(X,\theta))\right] = I(\theta)$$

$$\frac{l'(\theta_0)}{\sqrt{n}} = \frac{\sum \frac{d}{d\theta} log(f(X,\theta))}{\sqrt{n}} \xrightarrow{d} N(0, I(\theta))$$

$$\frac{l'(\theta_0)}{\sqrt{n}} \div \frac{l''(\theta_0)}{n} \xrightarrow{d} N\left(0, \frac{I(\theta)}{I(\theta)^2}\right) = N\left(0, \frac{1}{I(\theta)}\right), \text{ by Slutsky's theorem}$$

6.5 Bayes estimator

- **Prior distribution:** $\pi(\theta)$ the distribution of random variable Θ from which model parameter θ is drawn.
- **Conditional distribution:** $f(\{X_1, \ldots, X_n\} \mid \theta)$ is the conditional distribution of the data given $\Theta = \theta$
- **Posterior distribution:** $\pi(\theta \mid \{X_1, \ldots, X_n\})$ is the density of the random variable Θ given the observed data

$$\pi(\theta \mid \{X_1, \ldots, X_n\}) = \frac{f(\{X_1, \ldots, X_n\} \mid \theta)\pi(\theta)}{m(\{X_1, \ldots, X_n\})}, \text{ for } m(\{X_1, \ldots, X_n\}) = \int_{-\infty}^{\infty} f(\{X_1, \ldots, X_n\} \mid \theta)\pi(\theta)dx$$

The **Bayes Estimator** is calculated as $E[\pi(\theta \mid \{X_1, \ldots, X_n\})]$.
For recognizeable functions, we can back into the posterior distribution. If the prior and conditional distribution are not recognizeable, then we use numerical methods, like MCMC to approximate the posterior distribution

6.5.1 Example Bayes estimator method

$$X \sim Poisson(\theta), \theta \in [0,1] \qquad \pi(\theta) = exp(\theta)/(e-1)$$

$$\pi(\theta \mid X) \propto \frac{exp(-\theta)\theta^X}{X!} * \frac{exp(\theta)}{e-1}\mathbb{I}[\theta \in [0,1]] \propto \theta^X \mathbb{I}[\theta \in [0,1]]$$

$$\pi(\theta \mid X) = (X+1)\theta^X, \text{ observing } Beta(x+1,1) = \frac{\Gamma(x+2)}{\Gamma(X+1)\Gamma(1)}\theta^x = (x+1)\theta^x, \theta \in [0,1]$$

$$E[\pi(\theta \mid X)] = \int_0^1 \theta(X+1)\theta^X d\theta = \frac{X+1}{X+2}$$

6.5.2 Bayes estimator properties

All admissable estimators are Bayes Estimators. An estimator, $T'(\theta)$, is inadmissable if $\exists T$ such that

$$E_\theta[(T-\theta)^2] \leq E_\theta[(T'-\theta)^2] \; \forall \; \theta \text{ and } E_\theta[(T-\theta)^2] < E_\theta[(T'-\theta)^2] \text{ for some } \theta$$

6.6 Sufficiency

A test statistic, $T = T(X_1, \ldots, X_n)$ is **sufficient** for θ if $f(X_1, \ldots, X_n \mid T = t)$ does not depend on θ

The claim with a sufficient statistic is that there is no loss in throwing away the data as long as you keep the sufficient statistic

6.6.1 Fischer's Factorization Theorem

The **Fischer's Factorization Theorem** states that

$$T(X_1, \ldots, X_n) \text{ is sufficient for } \theta \iff \text{ joint density } f(X_1, \ldots, X_n, \theta) = g(T(X_1, \ldots, X_n), \theta)h(X_1, \ldots, X_n)$$

Proof: TODO

6.6.2 Rao-Blackwell Theorem

The **Rao-Blackwell Theorem** states
For $\hat{\theta}$ an estimator of θ with $E(\theta) < \infty$ and T sufficient with $\theta^* = E(\theta \mid T)$ then

$$E[(\theta^* - \theta)^2] \leq E[(\hat{\theta} - \theta)^2]$$

Proof: TODO

7 Hypothesis testing

- We assume data, $\{X_1, \ldots, X_n\}$ is generated by a distribution with parameter $\theta \in \Omega$ (could be a vector)
- The null hypothesis, H_0 and alternative hypothesis, H_1, are hypotheses for the true value of θ
  - A simple hypothesis is for a single value of θ, $H_i : \theta = \theta_i$
  - A composite hypothesis is for a range of θ, $H_i : \theta > 1$ or $H_i : \theta \neq \theta_0$
- The goal in testing is to construct a rule to decide whether to reject H_0
  - Want: $P_{H_0}(\text{falsely rejecting } H_0) = P_{H_0}(\text{Type I error}) \leq \alpha$
  - Want: maximal $P_{H_1}(\text{corectly rejecting } H_0) = 1 - P_{H_1}(\text{falsely accepting } H_0) = 1 - P_{H_1}(\text{Type II error})$
  - The rejection region, R, can be chosen to maximize correct rejections, subject to a Type I error constraint

7.1 Likelihood ratio

For simple hypotheses, the **Likelihood Ratio** is the ratio of the likelihoods under the alternative and null hypotheses. This ratio helps us boost correct rejections while limiting false rejections.

$$LR = \frac{f_{h_1}(\{X_1, \ldots, X_n\})}{f_{h_0}(\{X_1, \ldots, X_n\})}$$

We can define our rejection region, R using this the likelihood ratio. Specifically

$$R = \left\{ X : \frac{f_{h_1}(X)}{f_{h_0}(X)} \geq c \right\}$$

And constrain Type I error to level α by solving for c

$$P_{H_0}(\text{Type I error}) = P_{H_0}(R) = P_{H_0}\left(\frac{f_{h_1}(X)}{f_{h_0}(X)} \geq c \right) = \alpha$$

Our power then becomes $P_{H_1}(R)$

7.2 Neyman-Pearson lemma

For *simple hypotheses*, H_0, H_1, the **Neyman-Pearson lemma** states that the **Likelihood Ratio** level-α test, which rejects H_0 when $LR \geq c$, maximizes power, $P_{H_1}(LR \geq c)$. Any other level-α test, R', has $P_{H_1}(R') \leq P_{H_1}(LR \geq c)$
Proof:

Let $\phi(x) = \{1 \text{ if } x \in R; 0 \text{ otherwise}\}$, $\phi'(x) = \{1 \text{ if } x \in R'; 0 \text{ otherwise}\}$
Let $S^+ = \{x : \phi(x) = 1, \phi'(x) = 0\}$, $S^- = \{x : \phi(x) = 0, \phi'(x) = 1\}$

$\int_{-\infty}^{\infty} (\phi(x) - \phi'(x))(f_1(x) - cf_0(x))dx = \int_{S^+ \cup S^-} (\phi(x) - \phi'(x))(f_1(x) - cf_0(x))dx$, since 0 when $\phi(x) = \phi'(x)$

$\int_{S^+ \cup S^-} (\phi(x) - \phi'(x))(f_1(x) - cf_0(x))dx \geq 0$, since two differences are always opposing

$\int_{S^+ \cup S^-} (\phi(x) - \phi'(x))f_1(x)dx \geq \int_{S^+ \cup S^-} (\phi(x) - \phi'(x))cf_0(x)dx \geq 0$, since $RHS = c[\alpha - \alpha'] \geq 0$

$\int_{S^+ \cup S^-} \phi(x)f_1(x)dx \geq \int_{S^+ \cup S^-} \phi'(x)f_1(x)dx \iff P_{H_1}(R) \geq P_{h_1}(R')$

7.3 Uniformly Most powerful test (UMP)

The **Most Powerful** test is the test which maximizes power under simple hypotheses, H_0, H_1. The Neyman-Pearson Lemma tells us that the MP level-α test is the likelihood ratio test.

The **Universally Most Powerrful** test is the test that which maximizes power under composite hypotheses, H_1. That is, for $H_1 : \theta > a$ composite, the test is MP level-α for all simple $\tilde{H}_1 \in H_1$.

The general process for showing UMP is

- Consider simple hypotheses, H_0 vs. $\tilde{H}_1$
- Apply the Neyman-Pearson Lemma to find MP test for H_0 vs. $\tilde{H}_1$
- Show that the test doesn't depend on the choice $\theta_i \in H_1$

7.3.1 Example LR and UMP test

$X_i, \ldots, X_n \overset{i.i.d}{\sim} Poisson(\lambda), H_0 : \lambda = 1, H_1 : \lambda > 1$

$$LR(X) = \frac{\prod_{i=1}^n exp(-\lambda_1) * \frac{\lambda_1^{X_i}}{X_i!}}{\prod_{i=1}^n exp(-1) * \frac{1^{X_i}}{X_i!}} = \frac{exp(-n\lambda_1)\lambda_1^{\sum X_i}}{exp(-n)} = exp(n(1-\lambda_1))\lambda_1^{\sum X_i}, \text{ choosing some } \lambda_1 \in H_1$$

$$LR(X) \geq c \iff exp(n(1-\lambda_1))\lambda_1^{\sum X_i} \geq c \iff \lambda_1^{\sum X_i} \geq c' \iff \sum_{i=1}^n X_i \geq c''' = c$$

Under H_0, $\sum_{i=1}^n X_i \sim Poisson(n)$ and level-α test rejects when $\sum_{i=1}^n X_i \geq C_{n,1-\alpha}$ (upper $(1-\alpha)$ quantile of Poisson(n))

7.4 P-values

P-values answer "what is the smallest α that we would still reject H_0". For $T(X)$, a test statistic, and t, the statistic calculated from the data. Assume $T(X) \sim f_0(x)$, then $P_{H_0}[T(X) \geq t] = 1 - F_0(t)$

In the case $T(X) = \sqrt{n}\frac{\bar{X}_n}{\sigma} \sim N(0, \sigma)$ under H_0, then we have $P_{H_0}[\sqrt{n}\frac{\bar{X}_n}{\sigma} \geq t] = 1 - \Phi(t)$

7.5 Generalized Likelihood Ratio test

The **Generalized Likelihood Ratio test** provides us a way to compare composite hypotheses.

$$R_n = \frac{\max_{\theta \in \Omega_0 \cup \Omega_1} L_n(\theta)}{\max_{\theta \in \Omega_0} L_n(\theta)} = \frac{\hat{\theta}_{MLE}}{\max_{\theta \in \Omega_0} L_n(\theta)}$$

Twice the log of the Generalized Likelihood Ratio follows a χ^2_d distribution with $d = k - k_0$ degrees of freedom

$$2\log(R_n) \sim \chi^2_d, \text{ with } d = k - k_0$$

7.5.1 GLR example I

$$X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\theta, \sigma^2),\ H_0 : \theta = 0,\ H_1 : \theta \neq 0$$

$$R_n = \frac{L_n(\bar{X}_n)}{L_n(0)} = \exp\left(\frac{n\bar{X}_n^2}{2\sigma^2}\right)$$

$$2\log(R_n) = \frac{n\bar{X}_n^2}{2\sigma^2} = Z^2 \sim \chi^2_1, \text{ where } Z \sim N(0,1)$$

7.5.2 GLR example II

The **Poisson Dispersion Test**

$$X_1, \ldots, X_n \overset{i.i.d}{\sim} Poisson(\lambda_i),\ H_0 : \lambda_1 = \cdots = \lambda_n,\ H_1 : \text{ not } \lambda_i \text{ all equal}$$

$$R_n = \frac{L_n(\hat{\lambda}_{MLE_1}, \ldots, \hat{\lambda}_{MLE_n})}{L_n(\bar{X}_n)} = \prod_{i=1}^n \left(\frac{X_i}{\bar{X}_n}\right)^{X_i}$$

$$2\log(R_n) = 2\sum_{i=1}^n X_i \log\left(\frac{X_i}{\bar{X}_n}\right) \sim \chi^2_{n-1}$$

$$2\log(R_n) \approx \sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{\bar{X}_n}, \text{ using Taylor approximations}$$

7.5.3 Testing multinomial distributions

We can constructing the generalized likelihood ratio in the multinomial models as well

$$X_1, \ldots, X_n \sim multi(n, p_1, \ldots, p_n),\ H_0 : p_j = p_j(\theta)$$

Unrestrained MLE: $\hat{p}_j = \frac{X_j}{n}$, MLE under H_0: $\hat{\theta}_{MLE}$

$$R_n = \frac{\frac{n!}{X_1!\ldots X_n!}\hat{p_1}^{X_1}\ldots\hat{p}_n^{X_n}}{\frac{n!}{X_1!\ldots X_n!}p_1(\hat{\theta})^{X_1}\ldots p_n(\hat{\theta})^{X_n}} = \prod_{j=1}^{n}\left(\frac{\hat{p}_j}{p_j(\hat{\theta})}\right)^{X_j}$$

$$2\log(R_n) = 2\sum_{j=1}^{n} X_j \log\left(\frac{\hat{p}_j}{p_j(\hat{\theta})}\right) = 2\sum_{j=1}^{n} X_j \log\left(\frac{X_j}{np_j(\hat{\theta})}\right)$$

$$2\log(R_n) = 2\sum_{j=1}^{n} O_j \log\left(\frac{O_j}{E_j}\right), \text{ for } O_j = X_j \text{ and } E_j = np_j(\hat{\theta})$$

We can approximate this equality in Generalized Likelihood Ratios using the Taylor approximation to get the **Chi Squared Statistic**

$$2\log(R_n) = 2\sum_{j=1}^{n} O_j \log\left(\frac{O_j}{E_j}\right) \approx \sum_{j=1}^{n}\frac{(O_j - E_j)^2}{E_j} \sim \chi^2_d, \text{ with } d = k - k_0$$

In general, we can determine the degrees of freedom under H_0, k_0, as $(r-1)+(c-1)$ and under H_1, k, as $r*c-1$
The **Chi-square Test of Homogeneity** tests $H_0 : \pi_{i1} = \cdots = \pi_{iJ}$ with statistic

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(I-1)(J-1)}$$

8 Helpful applied methods

8.1 Derivation of Linear regression

TODO

8.2 Confidence intervals

TODO

8.3 Comparing two samples

TODO

8.4 Fischer's Exact test

TODO