Statistics

> The note of UIUC course STAT 400 in Fall 2023, by Zory Zhang. In case of any broken math renderring in Github preview, please open this markdown file using your own markdown editor. PDF version (maybe not up-to-date) can be found here.

For robust statistics, see the page.

Reminder

- The two properties of variance can save a lot of work.

Preliminaries

The following content is based on the course STAT400@UIUC. Probability knowledge is assumed.

Textbook

Probability and Statistical Inference, Tenth Edition; by Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman; Hoboken, NJ: Pearson; 2020.

Ch0 Matrix cookbook

- The matrix cookbook
- a for scalar, $\underline{a}$ for vector, A for matrix.
- X for random variable, $\underline{X}$ for random vector, $\mathbf{X}$ for random matrix.
- $\mathbb{E}[\underline{a}^T\underline{X}] = \underline{a}\cdot\mathbb{E}\underline{X}$; $\mathbb{E}[A\underline{X}] = A\mathbb{E}\underline{X}$; $\mathbb{E}A\mathbf{X}B = A\mathbb{E}[\mathbf{X}]B$.
- $\underline{\mu} := \mathbb{E}\underline{X}, \underline{v} := \mathbb{E}\underline{Y}$.

Ch1 Probability

1.1 Discrete distributions

> Def. (**Bernoulli distribution**, $X \sim Be(p)$) Bernoulli trial is that $A \in \mathcal{F}$, and we call the trial a success if A occurs. Bernoulli distribution is based on single Bernoulli trial.
> $m_1 = m_2 = p, Var = p(1-p)$.

> Def. (**Binomial distribution**, $Y \sim Bin(n,p)$) Perform n independent Bernoulli trials with $p = \mathbb{P}(A)$, and let Bernoulli r.v.s. $X_1, X_2 \dots X_n$ be the indicator funciton of success of the experiments. Let $Y := \sum X_i$, then the p.m.f. $f_Y(k) = \binom{n}{k}p^k(1-p)^{n-k}, k = 1, 2, \dots, n$. **Binomial process** is that Y_n to be the number of successes in the first n $Be(p)$ trials.

> Def. (**Geometric distribution**, $W \sim Geom(p)$) Keep performing Bernoulli trials until the first success, and let $W :=$ the waiting time, then $f_W(k) = (1-p)^{k-1}p, k = 1, 2, \dots, F(x) = 1-(1-p)^x$.

> Def. (**Negative Binomial distribution**, $W_r \sim NB(r,p)$) Let $W_r :=$ the waiting time for r successes, then $f_{W_r}(k) = \binom{k-1}{k-r}p^{r-1}(1-p)^{k-r}p, k = 1, 2, \dots$

> Def. (**Approximate poisson process** with param $\lambda > 0$) Let the number of **occurrences** of some event in a given **continuous interval** be counted. If
>
> 1. The number of occurrences in nonoverlapping subintervals are independent.
> 2. The probability of exactly one occurence in a sufficiently short subinterval of length h is approximately λh.
> 3. The probability of more than one occurence in a sufficiently short subinterval of length is essentially 0.

> Def. (**Poisson distribution**, $X \sim poisson(\lambda)$) An approximation of number of occurances in an interval of length 1 is given. Consider partition the unit interval into n subintervals. As n becomes larger, the length of subinterval become smaller and thus more likely the last two properties hold. For a large n, $\mathbb{P}(X = k) \sim \binom{n}{k}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k}$, and it converges to $f_X(k) := \frac{\lambda^k}{k!}e^{-\lambda}, k = 0, 1, 2, \dots$

Rmk. $M(t) = e^{\lambda(e^t-1)}$, then $\mu = \sigma^2 = \lambda$.

E.g. Number of customers of a shop between 5-6pm.

> Def. (**Hypergeometric distribution**) $f_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$. $\mathbb{E}X = nN_1/N, VarX = n\frac{N_1}{N}\frac{N-K}{N}\frac{N-n}{N-1}$. The sample

> Def. (**Multinomial**) Multi-outcome version of hypergeometric. $f_X(x) = \frac{n!}{\prod x_i!}\prod p_i^{x_i}$, where $\mathbb{E}X_i = np_i, VarX_i = np_i(1-p_i)$.

1.2 Continuous distributions

> Def. (**Exponential distribution**, $X \sim Exp(\lambda)$) Define X as the waiting time for the first occurance in a poisson process. Then $F_X(x) = 1 - \mathbb{P}(\text{no occurences in [0,x]})$, therefore
>
> $$F_X(x) = \begin{cases} 0, x < 0 \\ \\ 1 - e^{-\lambda x}, x \geq 0 \end{cases}, f_X(x) = F_X'(x) = \begin{cases} 0, x < 0 \\ \lambda e^{-\lambda x}, x \geq 0 \end{cases}$$

Rmk.

1. It's usually expressed as $\theta = \frac{1}{\lambda}, f(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}}$.
2. $M(t) = \frac{1}{1-\theta t}$ for $t < \frac{1}{\theta}$, thus $\mu = \theta, \sigma^2 = \theta^2$, Fisher's moment coefficient of skewness is 2.
3. (Memoryless) $\mathbb{P}(X > s + t | X > t) = \mathbb{P}(X > s)$.
4. (Failture/hazard rate for positive r.v.) $r(t) = \lim_{h \downarrow 0}\frac{1}{h}\mathbb{P}(X \leq t + h | X > t) = \frac{f(t)}{1-F(t)} = \lambda$.

> Def. (**Gamma function**) For $t > 0, \Gamma(t) = \int_0^\infty y^{t-1}e^{-y}\,\mathrm{d}y$.

Rmk. $\Gamma(t) = (-y^{t-1}e^{-y})_{y=0}^\infty + \int_0^\infty (t-1)y^{t-2}e^{-y}\,\mathrm{d}y = (t-1)\Gamma(t-1)$. $\Gamma(1) = 1$. When t is an integer, $\Gamma(t) = (t-1)!$.

> Def. (**Gamma distribution**, $X \sim Gamma(\alpha, \theta)$) For $x \geq 0$, $f_X(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha}x^{\alpha-1}e^{-x/\theta}$. In a poisson process, the waiting time of first occurence is $Exp(\theta = 1/\lambda)$, now gamma distribution describes the waiting time until α-th occurrence.

Rmk. $\mathbb{E}X = \alpha\theta, VarX = \alpha\theta^2$.

> Def. (**Normal/Gaussian distribution**, $X \sim \mathcal{N}(\mu, \sigma^2)$) $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$. Standard normal distribution $Y \sim \mathcal{N}(0,1)$, denote the p.d.f. as $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(\frac{-x^2}{2})$.

Rmk. To compute $I = \int_{-\infty}^{\infty} \phi(x)\ \mathrm{d}x$, consider

$$I^2 = \int_{-\infty}^{\infty} \phi(x)\ \mathrm{d}x \int_{-\infty}^{\infty} \phi(y)\ \mathrm{d}y$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(-\frac{x^2+y^2}{2})\ \mathrm{d}x\ \mathrm{d}y$$

$$= \frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty} \exp(-\frac{r^2}{2})r\ \mathrm{d}r\ \mathrm{d}\theta = 1$$

Rmk. For normal distribution, the i-th central moment of $N(\mu, \sigma^2)$ is $\sigma_i = \begin{cases} 0, & \text{if i odd,} \\ \sigma^i(i-1)!!, & \text{o.w.} \end{cases}$

> Def. (**(standard) Cauchy distribution**, $X \sim \mathcal{N}(\mu, \sigma^2)$) p.d.f. $f_X(x) = \frac{1}{\pi(1+x^2)}, F_X(x) = \arctan(x) + \frac{\pi}{2}$.

Rmk. (Heavy tail distribution without moment) Notice the $\mathbb{E}X = \int_{-\infty}^{\infty} |x| f_X(x)\ \mathrm{d}x = \infty$ does not exist.

1.3 Covariance

1. $\sigma_{XY} = Cov[X,Y] = \mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)] = \mathbb{E}(XY) - \mu_X\mu_Y$.
2. $V[\frac{X}{\sigma_X}] = 1$, $\rho_{XY} = Cov[\tilde{X}, \tilde{Y}] = \frac{Cov[X,Y]}{\sigma_X\sigma_Y} \in [-1,1]$.
3. The **covariance matrix** $V[\underline{X}] = \mathbb{E}[(\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu})^T] = \mathbb{E}\underline{X}\underline{X}^T - \underline{\mu}\underline{\mu}^T$.
4. The **variance of r.v.** in the form of $\underline{a}\cdot\underline{X}$: $V[\underline{a}\cdot\underline{X}] = \underline{a}^T V[\underline{X}]\underline{a}$; $V[A\underline{X}] = AV[\underline{X}]A^T$. When $||\underline{a}|| = 1$, it's taking the variance in the direction of $\underline{a}$, over the joint p.d.f. (consider variance as the width of the function picture along certain axis).

1.4 CLT

Prop. (**Independent normal**) $\sum_{i=1}^n c_i X_i \sim N(\sum c_i\mu_i, \sum c_i^2\sigma_i^2)$.

Cor. When all Xi are from the same distribution, $\frac{1}{n}\sum X_i \sim N(\mu, \frac{\sigma^2}{n})$.

Thm. (**CLT**) $\frac{1}{n}S_n := \frac{1}{n}\sum_{i=1}^n X_i \sim N(\mu, \frac{\sigma^2}{n})$.

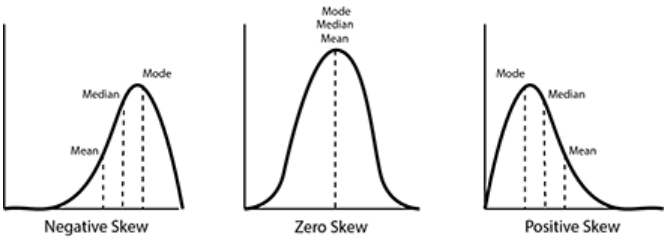
Rmk. Normal approximation work especially well on Binomial distribution when both $np, n(1-p) > 5$.

Rmk. (**Continuity correlation**) When normal approximation is applied to interger sample space, we need to adjust the event space. E.g. $\mathbb{P}(3 \le X \le 5) = \mathbb{P}(2.5 \le X \le 5.5)$.

1.5 Generating function

> Def. (**k-th moments**) $m_k := \mathbb{E}X^k = \int x^k f_X(x)\ \mathrm{d}x$.

Rmk. Third and forth moments are called skewness and kurtosis respectively.



Motivation. $G_X(s) = \mathbb{E}s^X, M_X(t) = \mathbb{E}e^{tX}, \phi_X(t) = \mathbb{E}e^{itX}$. Then $X \perp Y$, we have $G_{X+Y} = G_X G_Y$. The converse may not be true.

Rmk. $\frac{\mathrm{d}^k}{\mathrm{d}t^k} M_X(t)|_{t=0} = \int x^k \, \mathrm{d}F_X(x) = m_k$.

STAT

> Def. (**Sample mean, variance, and moment**) n samples $X_1, \dots X_n$, n observations $x_1, \dots x_n$.
>
> 1. The sample mean $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$;
> 2. The sample variance $s^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$;
> 3. The sample moment $\frac{1}{n} \sum_{i=1}^n x_i^k$.
>
> These sample statistical quantities are contrast to population statistical quantities. Essentially, they are just random variables before observed.

2.1 Point estimation

> Def. (**Point estimation**) Given a dist with some known params and a unknown param θ. The parameter space Ω is the range of all possible values of θ. Now we need to find our best guess for θ to have a good estimate of dist. Let an **estimator** of θ be $\hat{\theta} := u(X_1, X_2, \dots X_n)$. The instance $u(x_1, x_2, \dots x_n)$ is a **point estimate**.

Rmk. (**Maximum likelihood estimate, MLE**) For a **likelihood func** $L(\theta) := \prod_{i=1}^n f(x_i; \theta)$, the MLE $\hat{\theta} := \mathrm{argmax}_\theta L(\theta)$.

> Def. (**Bias of estimator**) Estimator $\hat{\theta}$ has bias $\mathbb{E}\hat{\theta} - \theta$. E.g. S_X^2 is an **unbiased estimator** of the variances.

Rmk. (**Method of moments, MOM**) Let i-th sample moment equal to i-th theoretical moment, i from 1 to the # unknown params, then solve the equations.

> Def. (**Chi-squared distribution**) If k iid $Z_i \sim N(0,1)$, then
> $\sum_{i=1}^k Z_i^2 \sim X^2(k) = Gamma(\alpha = k/2, \theta = 2)$. k is called the **degree of freedom**, DF.

Prop. $\frac{(n-1)S^2}{\sigma^2} \sim X^2(n-1)$.

Proof.

1. Assume

2.2 Confidence Interval, CI

> Def. (**Student's t-distribution**) $T := \frac{Z}{\sqrt{U/r}} \sim t_r$, where r is the DF, $Z \sim N(0,1), U \sim X^2(r), Z \perp U$. Then $f_T(t) = \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi r}\Gamma(\frac{r}{2})(1+t^2/r)^{(r+1)/2}}$.

Rmk. $\mathbb{E}T = 0, VarT = \frac{v}{v-2}$. When the sample size n is relatively big, the difference between normal and t-distribution is small.

E.g.

1. If the underlying distribution is normal, then $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1); T = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$. #NotCovered
2. The above can serve as approximation #NotCovered
3. "When the distribution is not normal but is unimodal (has only one mode), symmetric, and continuous, the approximation is usually quite good even for small n, such as n = 5. ... In almost all cases encountered in real applications, an n of at least 30 is usually adequate."
4. Usually n being larger than 30 is good enough. But if the underlying distribution is badly skewed or contaminated with occasional outliers, most statisticians would prefer to have a larger sample size—say, 50 or more—and even that might not produce good results.

> Def. (**Confidence interval, CI**) A $100(1-\alpha)\%$ confidence interval is a range believed with probability α for failing to contain the parameter of interest. The answer format: we are $\alpha\%$ confident that the mean number of xxx is between xxx and xxx.

E.g.

1. (**CI for μ when σ is known**) Let $z_{\frac{\alpha}{2}}$ be the $100(1-\frac{\alpha}{2})$-percentile of $N(0,1)$, then an **approximate 2-sided** $100(1-\alpha)\%$ CI for μ when σ is known is $\bar{x} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$, derived from $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ approximately and thus $\mathbb{P}(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}) \approx 1-\alpha$.
2. (**CI for μ when σ is unknown but n is large**) use s as the unbiased estimator of σ**, since $\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim N(0,1)$.
3. (**One sided CI**) Let z_α be the $100(1-\alpha)$-percentile of $N(0,1)$, then an approximate 1-sided CI for μ when σ is known can be $(\bar{x} - z_\alpha\frac{\sigma}{\sqrt{n}}, \infty)$, from $\mathbb{P}(Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_\alpha) \approx 1-\alpha$.
4. (**CI for μ when σ is unknown and n is small**) Let $t_{n-1,\frac{\alpha}{2}}$ be the $100(1-\frac{\alpha}{2})$-percentile of t_{n-1}, then when σ is unknown (need estimation s), it is $\bar{x} \pm t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$, derived from $\mathbb{P}(-t_{\frac{\alpha}{2}} \le T = \frac{\bar{X}-\mu}{s/\sqrt{n}} \le t_{\frac{\alpha}{2}}) \approx 1-\alpha$.
5. (**Sample size for target error when σ is known or n is large**) The question is to find a sample size so that we have high confident that the mean is within $\bar{X} \pm \epsilon$. In other word, the CI is within $\bar{X} \pm \epsilon$, where ϵ is called the **maximum error of the estimate**. Then $\epsilon := z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$. When we want to know what sample size n is needed for certain amount of error, we can solve n from that equation.
6. (**Sample size for target error when σ is unknown and n is small**) Note that $\mathbb{P}(\mu \in \bar{X} \pm t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}}) = 1-\alpha$, and $\mathbb{E}(t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}})^2 = t_{n-1,\frac{\alpha}{2}}\frac{\sigma^2}{\sqrt{n}}$. Now if we hope $\mathbb{E}\epsilon^2 := \mathbb{E}(t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}})^2 \le \epsilon_u^2$, then we want $\frac{n}{t^2_{n-1,\frac{\alpha}{2}}} > \frac{s^2}{\epsilon_u^2}$. We can approximate it by conducting a preliminary experiment first to have sample variance, then get a sense of the n by using (5), and then improve the gap one observation at a time until it obeys the inequality.

7. (**CI for $\mu_X - \mu_Y$ when common σ is unknown and n is small**) Suppose $X_1, \dots X_{n_X}$ and $Y_1, \dots Y_{n_Y}$ independently from two normal $N(\mu_X, \sigma^2), N(\mu_Y, \sigma^2)$.
   - $Z := \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}}} \sim N(0,1)$.
   - OTAH, $U := \frac{(n_X - 1)S_X^2}{\sigma^2} + \frac{(n_Y - 1)S_Y^2}{\sigma^2}$ is the sum of two independent chi-square r.v. thus $U \sim X^2(n_X + n_Y - 2)$.
   - The independence between the sample means and sample variances implies that $Z \perp U$. Thus $T := \frac{Z}{\sqrt{\frac{U}{n_X + n_Y - 2}}} \sim t_{n_X + n_Y - 2}$.
   - Then let $S_p := \sqrt{\frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}}, t_0 := t_{n_X + n_Y - 2, \alpha/2}$, we have the CI: $\bar{X} - \bar{Y} \pm t_0 S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$.
8. (**CI for proportion** p) Treat proportion as event success with probability p. Assume independent $X_1, \dots X_n \sim Bernoulli(p)$, unbiased MLE $\hat{p} = \frac{Y}{n} := \frac{\sum X_i}{n}$, where $Y \sim Bin(n, p)$.
   - By CLT, $\frac{\hat{p} - \mathbb{E}\hat{p}}{\sqrt{Var\hat{p}}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$ approximately when n is large enough.

$$\mathbb{P}(-z_{\frac{\alpha}{2}} \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

$$\iff \mathbb{P}\left(\frac{Y}{n} - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \frac{Y}{n} + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

   - Make additional approximation by replace p in end points with $\hat{p} := \frac{Y}{n}$.
9. (**CI for variance**)

2.3 Hypothesis testing

Multivariate normal

1. Standard normal: $\underline{Z} \sim N(\underline{0}, I)$.
2. (Stretch) $\underline{X} = D\underline{Z}$ where $D = diag(\sigma_1, \sigma_2, \dots)$, then $\underline{X} \sim N(\underline{0}, D^2)$.
3. (Rotate) $\underline{Y} = Q\underline{X}$ where $Q^TQ = I$ (orthogonal), then $\underline{Y} \sim N(\underline{0}, V)$. Given D to find V and Q, consider that by (Covariance-4),

$$V = V[Y] = V[QX] = QV[X]Q^T = QD^2Q^T$$

   Eigenvalue decomposition solves this problem. Say $eigen(V) = \lambda_1, \dots$, then $\lambda_i = \sigma_i^2$.

Hypothesis Testing

Kullback–Leibler divergence

a measure of how one probability distribution *P* is different from a second, reference probability distribution *Q*.