STAT 231 - S25
Statistics
3-page Cheat Sheet

With Prof James Huang

This is my final exam cheat sheet. They don't cover everything, but hopefully they are still helpful.

Josiah Plett

STAT 231 – Final Exam Cheat Sheet – Josiah Plett – https://plett.dev/notes

Unit: Individual measurable item.

Population: Static collection of units.

Process: Produced collection of units.

Variate: Any varying value.
↳ continuous, discrete, categorical, ordinal

Attribute: Function of a variate.

Skewness: <0 =0 >0

Kurtosis: 1.8 =3 =5

IQR: inter-quartile range $q(0.75)-q(0.25)=IQR$

	A	$\bar{A}$
B	y_{11}	y_{12}
$\bar{B}$	y_{21}	y_{22}

Relative Risk of A in B (vs $\bar{B}$): $RR=\frac{y_{11}/(y_{11}+y_{12})}{y_{21}/(y_{21}+y_{22})}$

5# Summary: $\{y_{(1)}, q(0.25), q(0.5), q(0.75), y_{(n)}\}$

Sample Variance:

$$S^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i-\bar{y})^2=\frac{1}{(n-1)}\sum_{y}\left(y^2 f_y-n\mu^2\right)$$

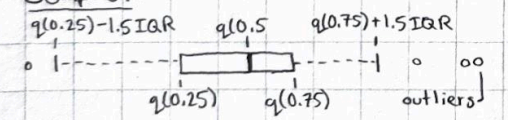
↑ σ^2

Covariance: $Cov(X,Y)=E[(X-\mu_X)(Y-\mu_Y)]$

Correlation: $corr(X,Y)=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$

Choose: $\binom{n}{x}=\frac{n!}{x!(n-x)!}$

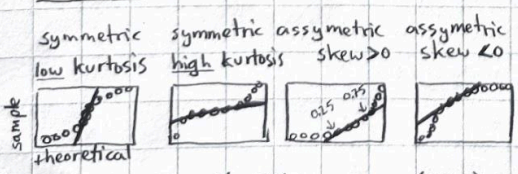
Boxplot:



Sample Correlation: $r=\frac{S_{XY}}{\sqrt{S_{XX}S_{XY}}}$

$$S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

QQPlot: Gaussian → linear plot



Central Limit Theorem (CLT):

$$Z_n=\frac{\sqrt{n}(\bar{Y}-\mu)}{\sigma}\quad Z_n\xrightarrow{n\to\infty}G(0,1)$$

$(Y_1,\dots,Y_n$ IID)

Maximum Likelihood Estimate (MLE):

$$L(\theta)=\prod_{i=1}^{n}f_Y(y_i;\theta)\qquad \hat{\theta}=\underset{\theta\in\Omega}{\operatorname{argmax}}\,L(\theta)\qquad R=\frac{L(\theta)}{L(\hat{\theta})}$$

Model	Params, PDF, E[x], Var[x]	MLE	$R(\theta)=\frac{L(\theta)}{L(\hat{\theta})}$	(G) Pivotal Quantity
Poisson	λ; $\frac{\lambda^x e^{-\lambda}}{x!}$; λ; λ	$\bar{y}$	$\left(\frac{\theta}{\hat{\theta}}\right)^{n\hat{\theta}}e^{n(\hat{\theta}-\theta)}$, $\theta>0$	$\frac{\tilde{\theta}-\theta}{\sqrt{\tilde{\theta}/n}}$
Binomial	n,p; $\binom{n}{x}p^x(1-p)^{n-x}$; np; $np(1-p)$	$\frac{y}{n}$	$\left(\frac{\theta}{\hat{\theta}}\right)^{y}\left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n-y}$, $0<\theta<1$	$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}}$
Geometric	p; $(1-p)^{x-1}p$; $1/p$; $\frac{1-p}{p^2}$	$\frac{1}{1+\bar{y}}$	$\left(\frac{\theta}{\hat{\theta}}\right)^{n}\left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n\bar{y}}$, $0<\theta<1$	
Exponential	λ; $\lambda e^{-\lambda x}$, $x\geq 0$; $1/\lambda$; $1/\lambda^2$	$\bar{y}$	$\left(\frac{\hat{\theta}}{\theta}\right)^{n}e^{n(1-\hat{\theta}/\theta)}$, $\theta>0$	$\frac{\tilde{\theta}-\theta}{\tilde{\theta}/\sqrt{n}}$
Gaussian (plain)	μ,σ; $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma}}$; μ; σ^2	$\hat{\mu}=\bar{y}$, $\hat{\sigma}=\sqrt{s^2}$	$e^{-\frac{n}{2\sigma^2}(\mu-\bar{x})^2}$	$\frac{\tilde{\theta}-\theta}{\sigma/\sqrt{n}}\Rightarrow\frac{\bar{x}-\mu}{s/\sqrt{n}}$
Simple Linear Regression	$Y_i\sim G(\alpha+\beta x_i,\sigma)$, $i=1,2,\dots n$ independent	Least Squares Estimates (MLE): $\hat{\alpha}=\bar{y}-\beta\bar{x}$; $\hat{\beta}=\frac{S_{xy}}{S_{xx}}=r\cdot\frac{S_y}{S_x}$; $\hat{\sigma}^2=\frac{1}{n}(S_{yy}-\hat{\beta}S_{xy})$		

	Unknown	Pivotal Quantity	100p% Confidence Interval
$G(\mu,\sigma)$, σ known	μ	$\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\sim G(0,1)$	$\bar{y}\pm a\sigma/\sqrt{n}$ $\quad a=qnorm(\frac{1+p}{2})$
$G(\mu,\sigma)$, σ unknown	μ	$\frac{\bar{Y}-\mu}{S/\sqrt{n}}\sim t(n-1)$	$\bar{y}\pm bs/\sqrt{n}$ $\quad b=qt(\frac{1+p}{2},n-1)$
$G(\mu,\sigma)$, μ,σ unknown	Y	$\frac{Y-\bar{Y}}{S\sqrt{1+1/n}}\sim t(n-1)$	$\bar{y}\pm bs\sqrt{1+1/n}$ $\quad b=$
$G(\mu,\sigma)$, μ unknown	$\sigma^{(2)}$	$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$	$\left[\sqrt{\frac{(n-1)s^{2(2)}}{d}},\sqrt{\frac{(n-1)s^{2(2)}}{c}}\right]$ c, d below

a: $P(Z\leq a)=\frac{1+p}{2}$ **b:** $P(T\leq b)=\frac{1+p}{2}$, $T\sim t(n-1)$

c, d: $P(W\leq c)=\frac{1-p}{2}=P(W>d)$, $W\sim\chi^2(n-1)$ $\quad c=qchisq(\frac{1-p}{2},n-1)$ $\quad d=qchisq(\frac{1+p}{2},n-1)$

Exponential(θ)	θ	$\frac{2n\bar{Y}}{\theta}\sim\chi^2(2n)$	$\left[\frac{2n\bar{y}}{d_1},\frac{2n\bar{y}}{c_1}\right]$	c_1,d_1: $n-1\to 2n$

Problem: Descriptive, Causative, Predictive

Plan (structure)

Data (collection)

Analysis

Conclusion

Surveys: Sample, observational, experimental

ERRORS: Target Population —Study error→ Study Population —Sample error→ Sample —Measurement error→ Measured Variate Values

$$SSE=\sum(Y_i-\hat{Y}_i)^2$$

$$S_e=\sqrt{\frac{SSE}{n-2}}$$

$$\hat{r}_i=y_i-\hat{\alpha}-\hat{\beta}x_i$$

$$\hat{r}_i^*=\hat{r}_i/S_e$$

p-value based on Gaussian distribution: (approximate)

$$2P\left(Z\geq\left|\frac{\text{pivotal quantity}}{}\right|\right),\quad Z\sim G(0,1)$$

$2\cdot(1-pnorm(d))$

↳ use $\hat{\theta}\pm Z\cdot\frac{1}{\sqrt{\text{pivotal quantity without numerator}}}$

Binomial: $H_0:\theta=\theta_0$ (P(success)) $\quad D=|y-n\theta_0|$

Poisson: $H_0:\theta=\theta_0$ → average per interval $\quad D=|y-t\theta_0|$

Exponential: $H_0:\theta=\theta_0$ → average wait time $\quad D=\frac{2n\bar{y}}{\theta_0}$

Binomial: Calculate sample size

$$n \geq \left(\frac{a}{m}\right)^2 \hat{\theta}(1-\hat{\theta})$$

$a = Z$: value from p confidence level.
$m = (U(y) - L(y))/2$: ½ confidence interval.

χ^2 Distribution

$$f_k(x) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}$$

k: degrees of freedom

$\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$
$\Gamma(n) = (n-1)\Gamma(n-1)$ for $n > 1$
$\Gamma(1/2) = \sqrt{\pi}$

$F_k \underset{k\to\infty}{\sim} G(k, 2k)$

$Y_1, \dots, Y_n \sim G(0,1) \Rightarrow \sum_{i=1}^{n} Y_i^2 \sim \chi^2(n)$

$F_2 \sim \text{Exp}(2)$

$W_1, \dots, W_n \sim \chi^2(k_i) \Rightarrow \sum W_i \sim \chi^2(\sum k_i)$

Likelihood Interval ⇔ Confidence Interval

$100p\%$ L.I. is defined as $\{\theta : R(\theta) \geq p\}$

$100p\%$ L.I. is an approx $100q\%$ C.I. where:
$q = P(W \leq -\ln p)$, $W \sim \chi^2$

Student t Distribution

Let $Z \sim G(0,1)$, $Y \sim \chi^2(k)$ be independent: $T = \frac{Z}{\sqrt{Y/k}} \sim t(k)$

Gaussian Confidence Intervals

$$\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1) \quad \frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t(n-1) \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Equal-Tailed $100p\%$ C.I.: $P(-a \leq \underbrace{\text{pivotal quantity}}_{\text{pivotal quantity}} \leq a) = p \rightarrow \left[\bar{y} \pm a\sigma/\underbrace{\sqrt{n}}_{\text{denominator}}\right]$

One-sided $100p\%$ C.I.: $P(-a \leq \underbrace{p.q.}_{\text{upper}}) = p$ or $P(\underbrace{p.q.}_{\text{lower}} \leq a) = p$

Simple Linear Regression Intervals

$\hat{\beta} \pm a\, Se/\sqrt{S_{xx}}$
$\hat{\alpha} \pm a\, Se\sqrt{1/n + \bar{x}^2/S_{xx}}$
$\hat{\mu} \pm a\, Se\sqrt{1/n + (x-\bar{x})^2/S_{xx}}$

a: $P(T \leq a) = \frac{1+p}{2}$ where $T \sim t(n-2)$

↳ Constant variance, Independence, Normality, Linearity

Residual Plots

standardized residual is uncorrelated with x

Simple Linear Regression Question

Given: $Y_i = \alpha + \beta x_i + R_i$, $R_i \sim G(0,\sigma)$, $i = 1, \dots, n$ independently

Prediction Interval > Confidence Interval

a) Show $E[\tilde{\alpha}] = \alpha$ (note $E[\tilde{\beta}] = \beta$)

$E[\tilde{\alpha}] = E[\bar{Y} - \tilde{\beta}\bar{x}]$ (definition of α)
$= \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i) - \bar{x}\beta = \alpha$

d) Give 95% C.I. for β

$\hat{\beta} \pm c\, SE(\hat{\beta})$ where $c = qt(0.975, n-2)$

Two-variate Analysis

$H_0: P(x \cap y) = P(x) \cdot P(y)$

① Table is averages of real totals:

	high	low	
tall	20	5	25
short	15	10	25
	35	15	

⇨ same

	high	low	
tall	17.5	7.5	25
short	17.5	7.5	25
	35	15	

② $\lambda = 2\sum_{j=1}^{4} y_j \ln\left(\frac{y_j}{e_j}\right)$ y_j = observed, e_j = expected

③ p-value $= P(\chi_1^2 \geq \lambda)$ $(H_0: \lambda \sim \chi_1^2)$

Multinomial

Test $H_0: \mu_1 = \mu_2 \dots$ assuming same s.d.:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

evaluated

→ p-value $= 2 \cdot (1 - P(T \leq t))$

Pearson's χ^2 Goodness of Fit (dice)

$$D = \sum_{j=1}^{6} \frac{(Y_j - E_j)^2}{E_j} \sim \chi^2(6-1) \text{ under } H_0$$

d = input the values.

p-value $= P(D \geq d \mid H_0) \approx P(W \geq d)$, $W \sim \chi^2(5)$

Expected Frequency

$e_{\#} = n(\hat{\theta})e^{-\hat{\theta}}$ ← Poisson

Graphical Summaries

(Q1) $\bar{y}=1.58$ $s^2=\frac{\sum(y_i-\bar{y})^2}{n-1}=1.30$

$\sum_{i=1}^{n} y_i^2=(n-1)s^2+n\bar{y}=378.34$

t, χ^2 Properties

$Z_1,\dots,Z_{n+1}$ are IID $\sim G(0,1)$

$\chi^2(n)\sim Z_1^2+\dots+Z_n^2$

$t(n)\sim \frac{Z}{\sqrt{U/n}}=\frac{Z_{n+1}}{\sqrt{\sum Z_i^2/n}}$

$G(0,n)\sim nZ_1$

Likelihood Interval

40% likelihood interval → C.I.

a: $e^{-a^2/2}=0.4 \rightarrow a=\sqrt{-2\log(0.4)}$

q: $q=2\underbrace{P(Z\le a)}_{\Phi}-1 \rightarrow 100q\%$ C.I.

MLE

Given: $f_Y(y;\lambda)=\frac{e^{-\lambda}\lambda^y}{y!}$

STEPS:

1. $L(\lambda)=\prod_{i=1}^{n}\frac{e^{-\lambda}\lambda^{y_i}}{y!}=\text{constant}\cdot e^{-n\lambda}\lambda^{n\bar{y}}$
2. $\ell(\lambda)=\ln(L(\lambda))=\text{constant}-n\lambda+n\bar{y}\ln(\lambda)$
3. Solve $\ell'(\lambda)=0 \rightarrow \ell'(\lambda)=-n+\frac{n\bar{y}}{\lambda}$

General 100p% C.I.

1. $P(a\le \underbrace{\text{pivotal quantity}}\le b)=p$

   $P(W\le a)=P(W\ge b)=\frac{1-p}{2}$ where $W\sim$ table
2. Rearrange!

Approx n based on C.I.

1. ℓ = range of 100p% C.I.
2. ℓ should be $2\cdot(a\cdot\underbrace{\sigma/\sqrt{n}}_{\text{denominator of p.q.}})$

Hypothesis Test (G)

Given: $\bar{y}=101.7$ $s=13.5$ $Y_i\sim G(\mu,\sigma)$

Q: Run a test to determine if average could be 105.

1. $H_0: \mu=105$ (μ_0, base case) $H_1: \mu\neq 105$
2. D: $\frac{|\bar{Y}-\mu_0|}{S/\sqrt{n}}$
3. $2P(T\ge D)=2P(T\ge 0.843)\approx 0.417$

   (T: student t since we have S not σ; n-1 d.f.)

BONUS: 95% C.I.: $\bar{y}\pm t_{0.025,11}\cdot S/\sqrt{n}$ (qt)

"Two-sided" hypothesis test

Testing σ^2, with $Y_i\sim G(\mu,\sigma)$

$D=\frac{(n-1)s^2}{\sigma_0^2} \rightarrow \min\{2P(U\le D), 2P(U\ge D)\}$

Assymptotic Pivotal Q - Proof

Given: 100 coinflips, 58 heads.

Assume $Y\sim$ Binomial$(100,\theta)$, recall $E[Y]=n\theta$, $\text{Var}(Y)=n\theta(1-\theta)$, $\hat{\theta}=Y/n$ ($\theta=\frac{\text{heads}}{n}$)

Prove $Z_n=\frac{\hat{\theta}-\theta}{\sqrt{\theta(1-\theta)}\sqrt{n}}$ is asymptotic p.q. for θ.

1. $E[\hat{\theta}]=E[\frac{Y}{n}]=\frac{1}{n}E[Y]=\theta$
2. $\text{Var}[\hat{\theta}]=\text{Var}[\frac{Y}{n}]=\frac{1}{n^2}\text{Var}(Y)=\frac{\theta(1-\theta)}{n}$
3. Note $Y=\sum^{100}X_i$, where X_i IID $\sim$ Bernoulli(θ)
4. By CLT, big n: $Z_n=\frac{\sqrt{n}(\bar{x}-E[X_i])}{\sqrt{\text{Var}[X_i]}}\sim G(0,1)$
5. Since we know distribution of Z_n, it's a p.q.

Linear Regression Model

$\hat{\alpha}$: average y when $x=0$.

$\hat{\beta}$: mean increase in y per x.

Test linear relationship ($\hat{\beta}=0$)

$H_0: \beta=0$ $H_1: \beta\neq 0$

$t: \frac{|\hat{\beta}-\beta_0|}{S_e/\sqrt{S_{xx}}}=7.037$ (pt) d.f.: n-2

p-value: $2P(T\ge 7.037)=0.0005$

95% Confidence Interval for $\hat{\mu}$ (given x)

$\hat{\mu}(12)=\hat{\alpha}+\hat{\beta}x=2285.324$

★ $(\hat{\alpha}+\hat{\beta}x)\pm a\,Se\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$ ★

Likelihood Ratio Test Statistic TEST

Given: $n=25$ Poisson(θ) → $\hat{\theta}=\bar{y}$, $L(\theta)\propto\theta^{n\bar{y}}e^{-n\theta}$, $\bar{y}=3.16$

Use $H_0: \theta=2.5$, $H_1: \theta\neq 2.5$

1. $\Lambda(\mu_0)=-2\ln\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right)=-2\ln(R(\theta))=4.016$ ★
2. p-value $\sim P(W\ge 4.016)$, $W\sim\chi^2(1)$ (pchisq)

↑ I should've had more examples of this!!!!!!

R: TRUE: $W\le d$ FALSE: $W\ge d$