Agenda

1. Bootstrap Review and Confidence intervals

2. Data Visualization

3. Sufficiency and Rao-Blackwell

Boot strap

We have some estimator and want to construct CI.

What if it's hard to get mathematical results about estimators.

$\Rightarrow$ Bootstrap.

Use existing data generate "new" datasets by resampling our dataset to analyze distribution of estimator.

CI's

$\hat{\theta}$ point estimate

Normal interval

$$\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \sqrt{V_{BOOT}}$$

standard normal quartile

sample variance of bootstrap samples

Pivotal Interval

$$\left(2\hat{\theta} - \hat{\theta}^*_{1-\frac{\alpha}{2}} \, , \; 2\hat{\theta} - \hat{\theta}^*_{\frac{\alpha}{2}} \right)$$

bootstrapped sample quantiles

Percentile Interval

$$\left(\hat{\theta}^*_{\frac{\alpha}{2}},\ \hat{\theta}^*_{1-\frac{\alpha}{2}}\right)$$

bootstrapped sample quantiles

Assumptions

1. Normal CI only works if estimator normal.

2. Percentile interval Assume there exists function m s.t. $m(\hat{\theta})$ is normal, with variance ind of θ.

3. Pivotal interval NO ASSUMPTIONS

Exact when $\hat{\theta} - \theta$ is pivot.

Def. Pivot.

Function of estimator and parameter whose distribution does not depend on parameter.

dist of $\hat{\theta} - \theta$ does not depend on θ.

e.g. $X_1, \ldots, X_n \sim N(\mu, 1)$

Use estimator

$$\frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$$

$$\bar{X} \sim N(\mu, \tfrac{1}{n})$$

$$\bar{X} - \mu \sim N(0, \tfrac{1}{n})$$

Data Visualization

1. Histograms
2. Boxplots
3. QQPlots.

Quantile Review

Theoretical:

The quantile is the inverse

CDF.

$X \sim F$

$F(z) = P(X \leq z) = p$

$F^{-1}(p) = z$ ← quantile function.

<u>eg.</u> 0.975 quantile for $N(0,1)$

approx 1.96.

<u>SAMPLE:</u>

Suppose $X \sim F$

then $F(X) \sim \text{unif}(0,1)$

$$P\left(F(X) \leq z\right) = z$$

$X_{(k)}$ is kth order stat

$$F\left(X_{(k)}\right) \simeq \frac{k}{n}.$$

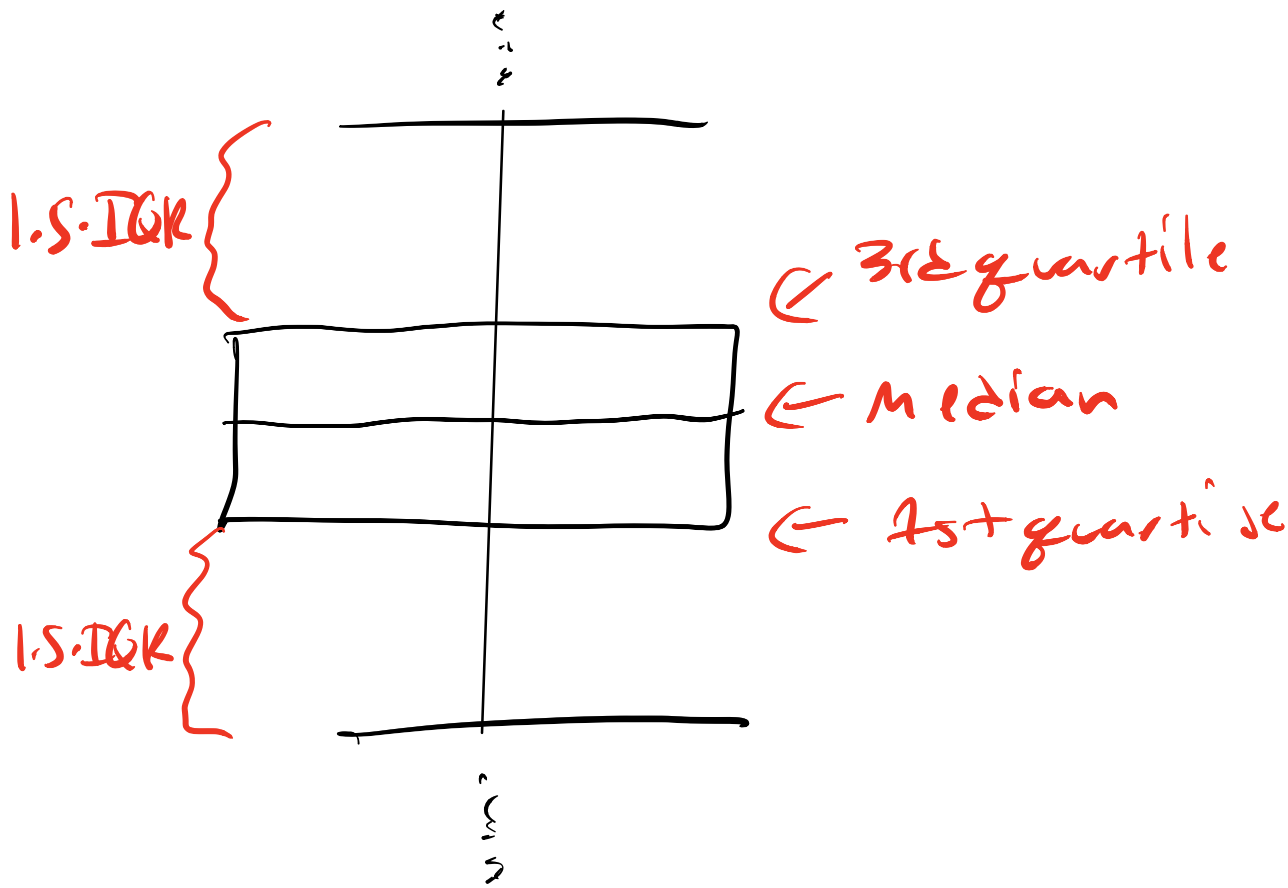
or

$$X_{(k)} \approx F^{-1}\left(\frac{k}{n}\right)$$

in practice

$$X_{(k)} \approx F^{-1}\left(\frac{k}{n+1}\right)$$

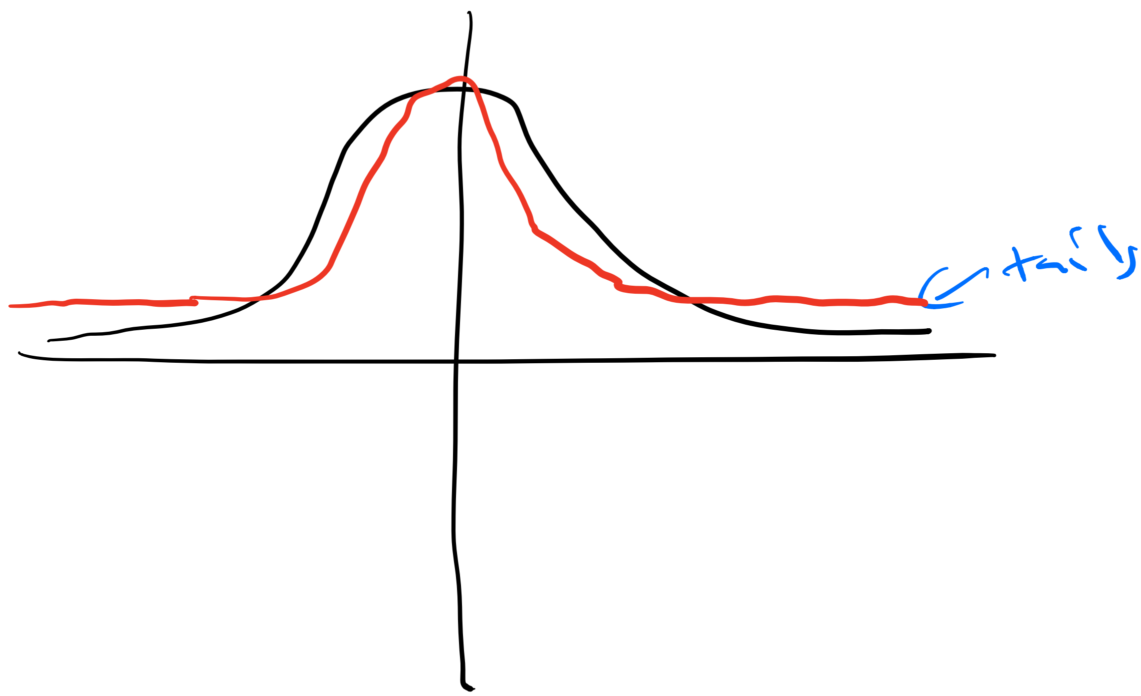
Boxplots



$$IQR = Q_3 - Q_1 = q_n(0.75) - q_n(0.25)$$

q_n, sample quantile.

Distribution tails

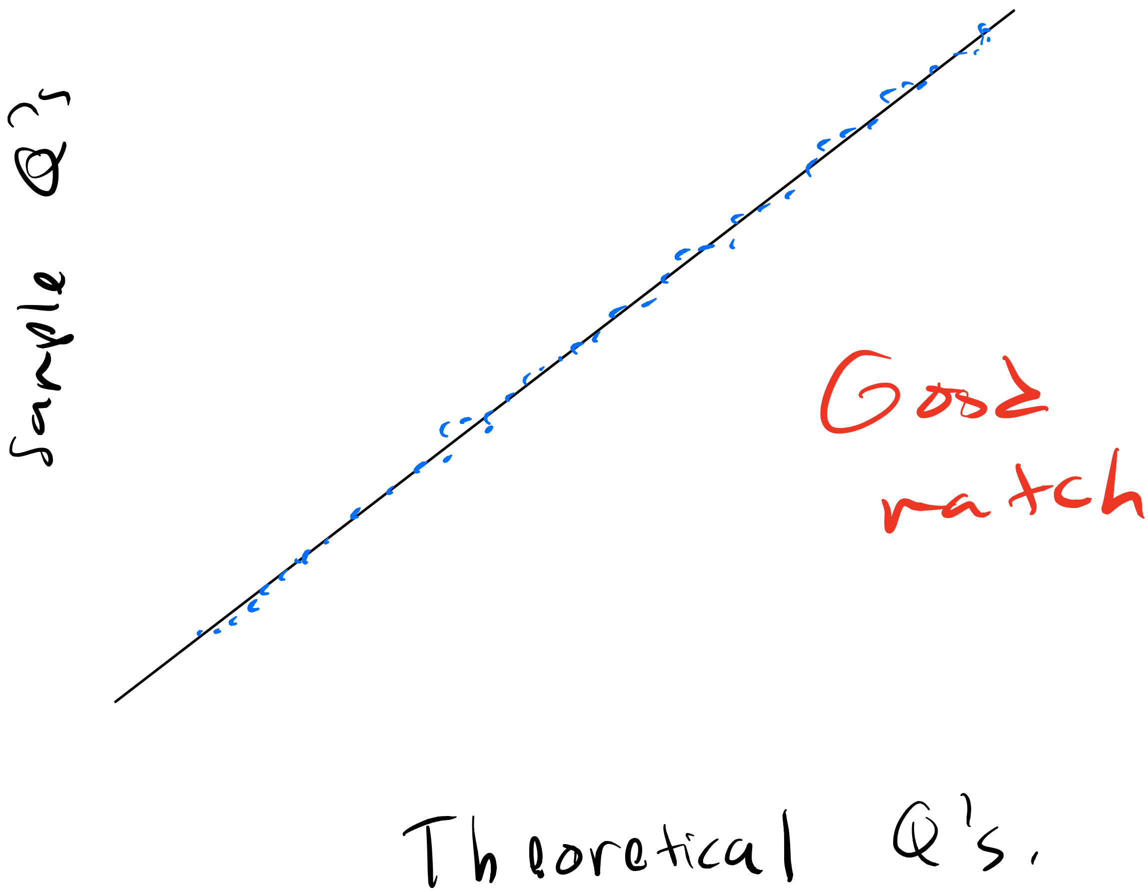


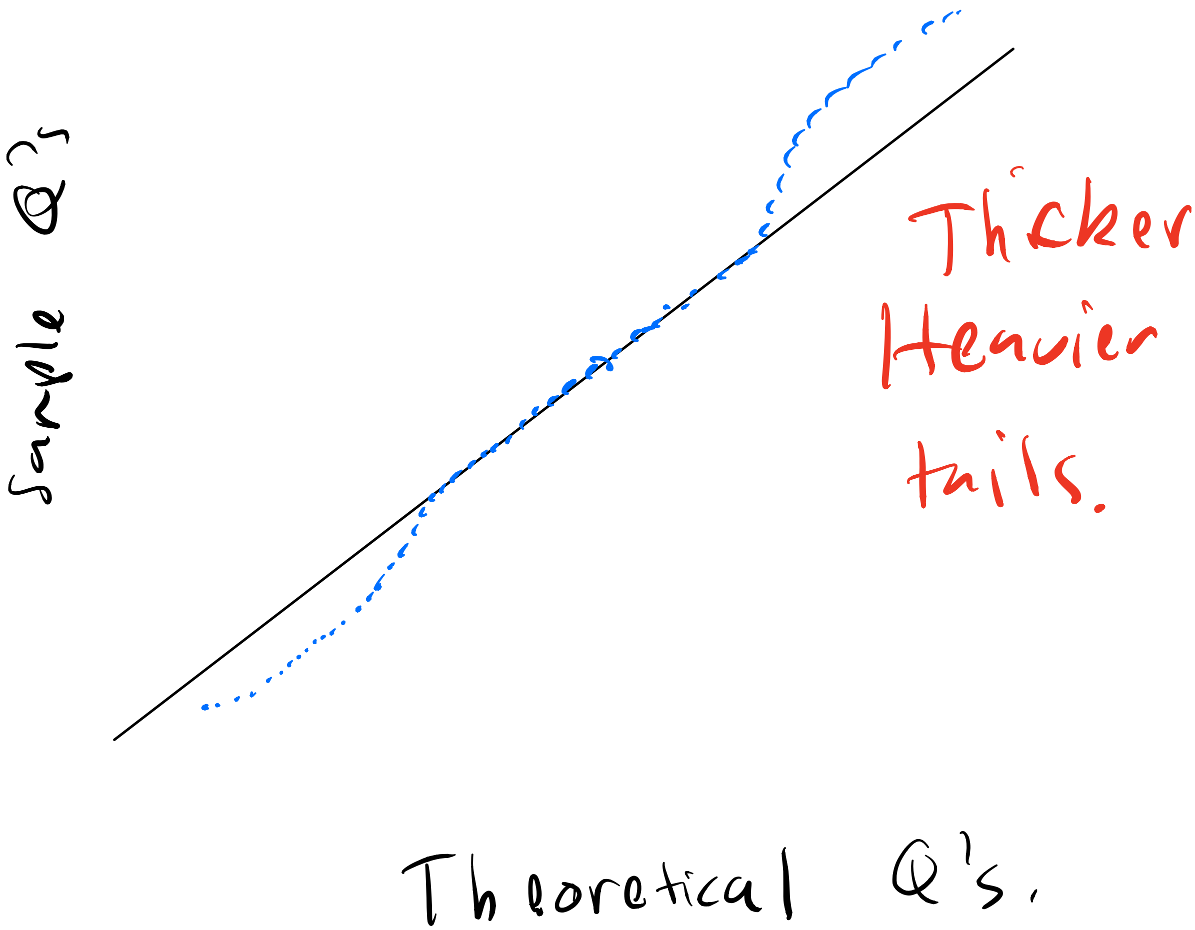
Red dist is thicker/heavier tailed than black distribution

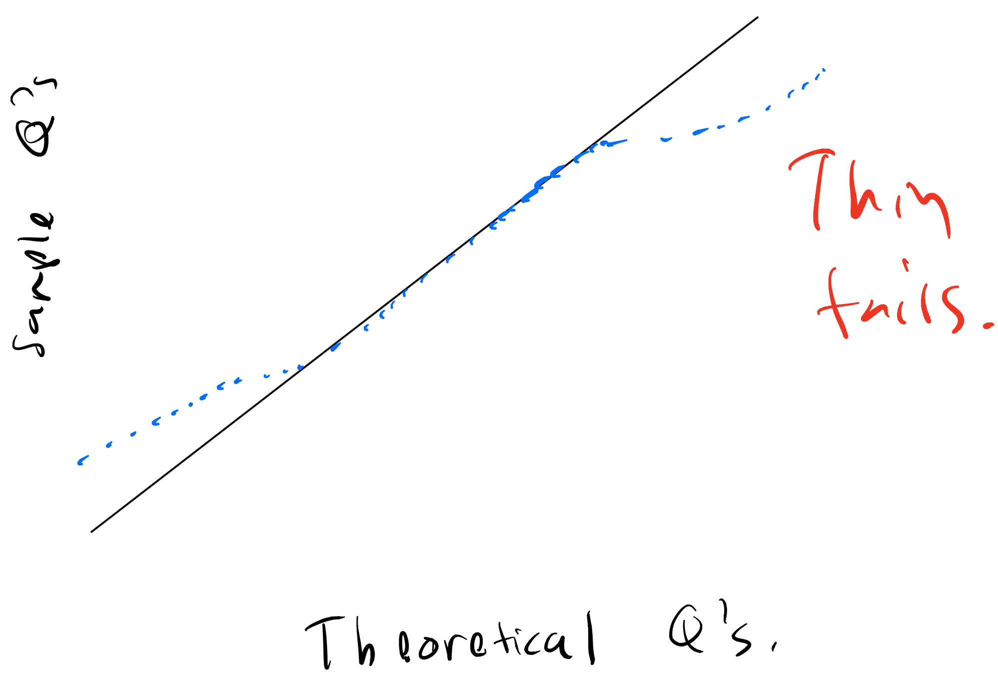
Ex. Thick tailed are t-distributions.

QQ-Plots

Suppose I believe data is normally distributed

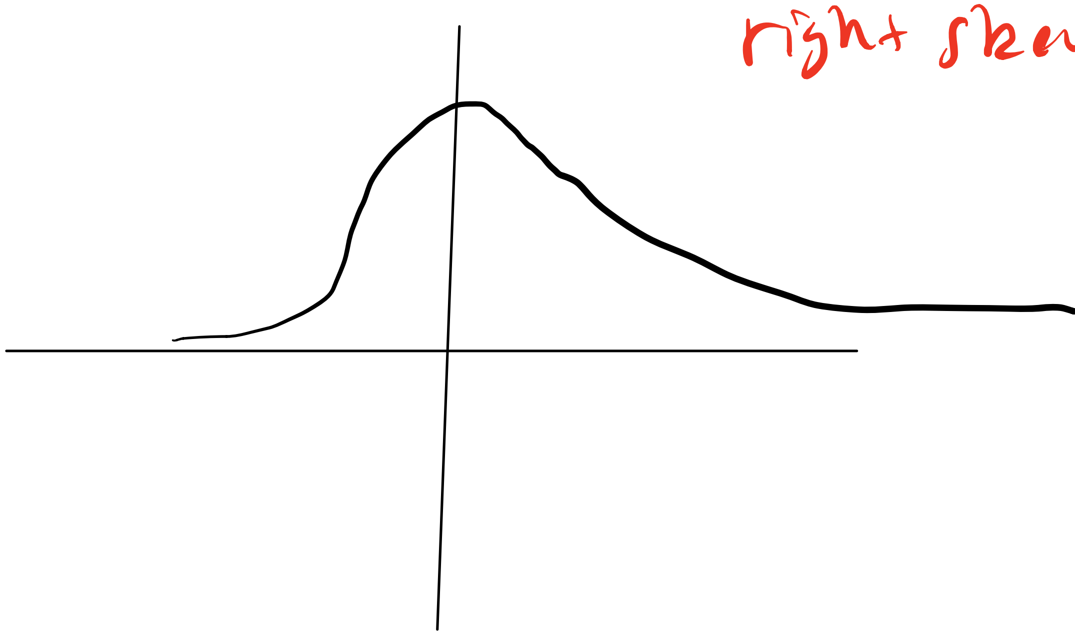



Sample Q's
Thicker
Heavier
tails.
Theoretical Q's.

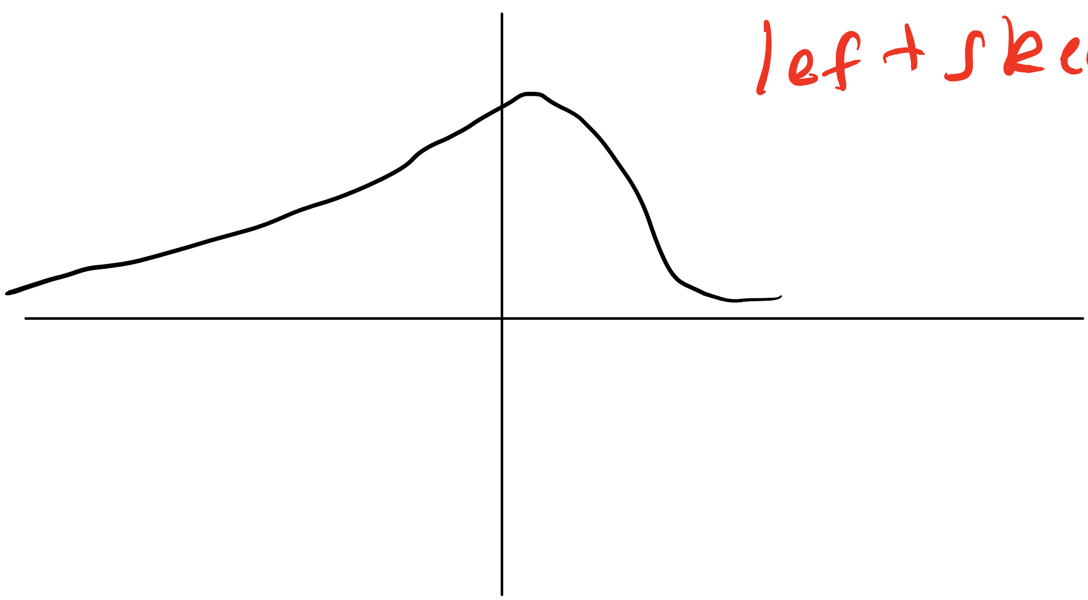

Sample Q's
Thin
tails.
Theoretical Q's.

Skewed Dist

right skew



left skew.



Sufficiency

Remember 4 nice things about estimator

1. Unbiased
2. Consistent
3. Efficient
4. Sufficient.

Data is costly to use and store. we only want we need for estimation procedure.

Consider estimating θ.

Can we find some function

$T(X_1, \ldots, X_n)$ that contains all info on θ?

Def. Sufficiency.

We say T is sufficient for θ if

$$X_1, \ldots, X_n \mid T(X_1, \ldots, X_n) = t$$

does not depend on θ for any t.

T is called a sufficient statistic.

You can prove this the long way.

The whole dataset is a sufficient statistic

Factorization Theorem

T is sufficient for θ if and only if

$$f_\theta(x_1,\dots,x_n) = \underbrace{g_\theta(T)}_{\text{solely a function of } T \text{ and } \theta.}\ \underbrace{h(x_1,\dots,x_n)}_{\text{no dep on } \theta!}$$

"θ depends on the data only through the sufficient statistic"

Ex. $X_1, \dots, X_n \overset{ind}{\sim} N(\mu, 1)$

Let's find suff. stat for μ (Fctrizn thm)

$$f_\mu(x_1, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i-\mu)^2\right)$$

$$= \frac{1}{(\sqrt{2\pi})^n} \exp\left(-\frac{1}{2} \sum (x_i-\mu)^2\right)$$

$$= \frac{1}{(\sqrt{2\pi})^n} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (x_i^2 - 2x_i\mu + \mu^2)\right)$$

$$= \frac{1}{(\sqrt{2\pi})^n} \exp\left(\frac{-1}{2}\sum_{i=1}^{n} x_i^2\right) \cdot \exp\left(\mu \sum_{i=1}^{n} x_i\right) \cdot \exp\left(\frac{-1}{2}\sum_{i=1}^{n} \mu^2\right)$$

$$= \overbrace{\frac{1}{(\sqrt{2\pi})^n} \exp\left(\frac{-1}{2}\sum_{i=1}^{n} x_i^2\right)}^{h(x_1,\ldots,x_n)} \cdot \underbrace{\exp\left(\mu \cdot n \cdot \bar{x}\right) \cdot \exp\left(\frac{-1}{2}\sum_{i=1}^{n} \mu^2\right)}_{g_\mu(\bar{x})}$$

$\bar{X}$ is a sufficient stat for μ.

note $\sum x_i$ is also sufficient.

Important Fact

If T is sufficient for θ, $\hat{\theta}_{MLE}$ is a function of T.

Why?

T sufficient if and only if

$$f_\theta(x_1, \dots, x_n) = g_\theta(T) \cdot h(x_1, \dots, x_n)$$

$$\log f_\theta(x_1, \dots, x_n) = \log\left(g_\theta(T)\right) + \log\left(h(x_1, \dots, x_n)\right)$$

maximizing this with respect to

θ, equivalent to maximizing

$\log(g_\theta(T))$

b/c h is a constant in terms of θ.

$\Rightarrow$ Sol to MLE is a function of T.

How can we measure goodness of estimator?

One way is $MSE(\hat{\theta})$

the mean squared error

$$MSE(\hat{\theta}) = E_\theta\left(\left(\hat{\theta}-\theta\right)^2\right)$$

"on average how far is estimator from true value squared"

Note we can decompose

$$MSE(\hat{\theta}) = \text{bias}^2 + \text{var}(\hat{\theta})$$
$$= \left(E(\hat{\theta})-\theta\right)^2 + \text{var}(\hat{\theta})$$

"bias-variance tradeoff"

Rao-Blackwell Theorem

Let $\hat{\theta}$ e estimator for θ. Assume $E(\hat{\theta}^2) < \infty$.

Suppose T sufficient for θ.

Let $\tilde{\theta} = E(\hat{\theta} \mid T)$

"Rao-Blackwellized estimator"

then $MSE(\tilde{\theta}) \leq MSE(\hat{\theta})$

Strict if $\tilde{\theta} \neq \hat{\theta}$

Note. $\tilde{\theta} = E(\hat{\theta} \mid T)$ has no dependence on θ because $\hat{\theta} \mid T$ has no dependence on θ, because T is sufficient for θ.

68. c) from Ch 8

$T = \sum_{i=1}^{n} X_i$ is sufficient

for λ, $X_i \overset{iid}{\sim} \text{pois}(\lambda)$

$$f_\lambda(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

$$= \underbrace{\lambda^{\sum x_i} \cdot e^{-n\lambda}}_{g_\lambda(\sum x_i)} \cdot \underbrace{\frac{1}{\prod_{i=1}^{n} x_i!}}_{h}$$

$$f_\lambda(x_1, \ldots, x_n) = g_\lambda(T)\, h(x_1, \ldots, x_n)$$

$$\lambda^{n\bar{x}} e^{-n\lambda} \frac{1}{\prod_{i=1}^{n} x_i!}$$