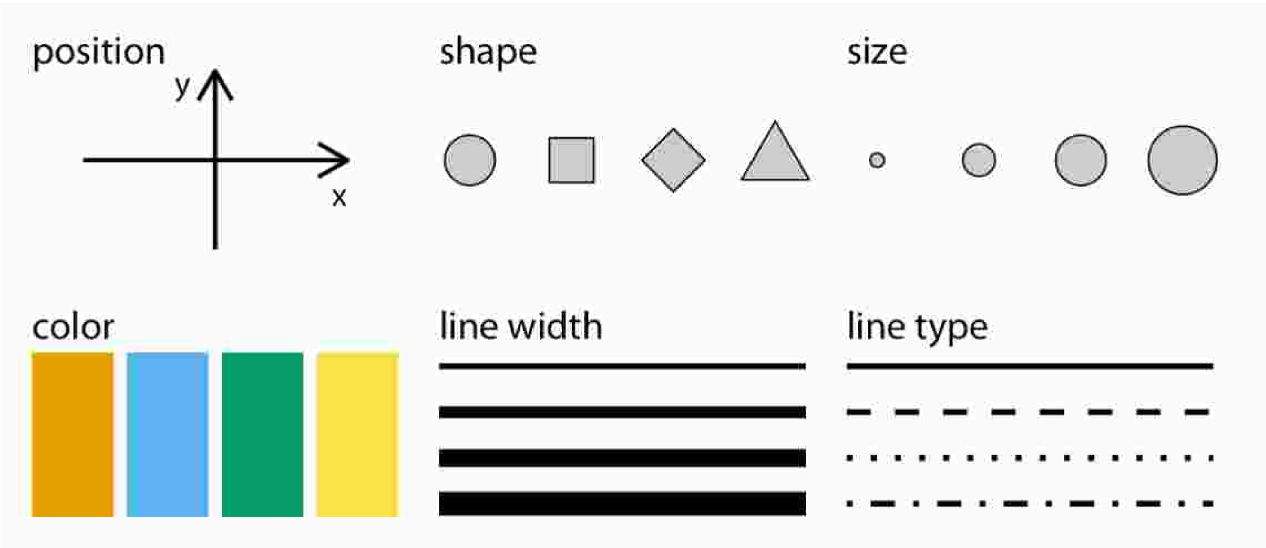
{EPIB607 CHEAT SHEET}

SAHIR RAI BHATNAGAR
(MCGILL UNIVERSITY)

DATA VISUALIZATION

- Commonly used aesthetics in data visualization: position, shape, size, color, line width, line type. Some of these aesthetics can represent both continuous and discrete data (position, size, line width, color) while others can only represent discrete data (shape, line type)
- Cynthia Brewer palettes: Sequential, Diverging, Qualitative



SAMPLING DISTRIBUTIONS, CLT, CI, P-VALUE

- **Standard error (SE)** of the sample mean is $\sigma/\sqrt{n}$
- **SE($\bar{y}$)** describes how far $\bar{y}$ could (typically) deviate from μ
- **SD(y)** describes how far an individual y (typically) deviates from μ (or from $\bar{y}$).

- **Paramter**: An unknown numerical constant pertaining to a population/universe, or in a statistical model. μ: population mean, π: population proportion, λ: population rate
- **Statistic**: A numerical quantity calculated from a sample. The empirical counterpart of the parameter, used to *estimate* it. $\bar{y}$: sample mean, p: sample proportion, $\hat{\lambda}$: sample rate

- **Sampling Distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population. The standard deviation of a sampling distribution is called a standard error.

ONE SAMPLE MEAN

σ	known	unknown
Data	$\{y_1, y_2, ..., y_n\}$	$\{y_1, y_2, ..., y_n\}$
Pop'n param	μ	μ
Estimator	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
SD	σ	$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$
SEM	$\sigma/\sqrt{n}$	$s/\sqrt{n}$
$(1-\alpha)100\%$ CI	$\bar{y} \pm z^{\star}_{1-\alpha/2}$(SEM)	$\bar{y} \pm t^{\star}_{1-\alpha/2,(n-1)}$(SEM)
test statistic	$\frac{\bar{y}-\mu_0}{\text{SEM}} \sim \mathcal{N}(0,1)$	$\frac{\bar{y}-\mu_0}{\text{SEM}} \sim t_{(n-1)}$

R CODE

Normal Distribution $Y \sim \mathcal{N}(\mu, \sigma^2)$

Cumulative Probabilities: `pnorm(q = , mean = , sd = , lower.tail = TRUE)`
Quantiles: `qnorm(p = , mean = , sd = )`
Linear regression: `lm(y ~ x, data = df)`
log link →: `glm(y ~ x, data = df, family=gaussian(link="log"))`

t Distribution $Y \sim t_{(df)}$

Cumulative Probabilities: `pt(q = , df = , lower.tail = TRUE)`
Quantiles: `qt(p = , df = )`

Binomial Distribution $Y \sim Binomial(N, \pi)$

Cumulative Probabilities $P(Y \leq k)$: `pbinom(q=,size=,prob=,lower.tail=TRUE)`
Cumulative Probabilities $P(Y > k)$: `pbinom(q=,size=, prob=,lower.tail=FALSE)`
Quantiles: `qbinom(p = , size = , prob = )`
Probabilities $P(Y = k)$: `dbinom(x = ,size = ,prob = )`
Logistic regression: `glm(y ~ x, data = df, family = binomial(link="logit"))`
log link →: `glm(y ~ x, data = df, family = binomial(link="log"))`

Poisson Distribution $Y \sim Poisson(\lambda)$

Cumulative Probabilities $P(Y \leq k)$: `ppois(q = , lambda = , lower.tail=TRUE)`
Cumulative Probabilities $P(Y > k)$: `ppois(q = , lambda =, lower.tail=FALSE)`
Quantiles: `qpois(p = , size = , prob = )`
Probabilities $P(Y = k)$: `dpois(x = ,lambda = )`
Poisson regression: `glm(y~x+offset(log(PT)),data=df,family=poisson(link="log"))`
identity link: `glm(y ~ -1 + PT + PT:x,family=poisson(link="identity"))`

ASSUMPTIONS

	z	t	Bootstrap
SRS	✓	✓	✓
Normal population	✓*	✓*	✗
needs CLT	✓*	✓*	✗
σ known	✓	✗	✗
Sampling dist. center at	μ	μ	$\bar{y}$
SD	σ	s	s
SEM	$\sigma/\sqrt{n}$	$s/\sqrt{n}$	SD(bootstrap statistics)

[a]*If population is Normal then CLT is not needed. If population is not Normal then CLT is needed.