richardwu.ca

STAT 231 Final Exam Guide

Introduction to Statistics

Surya Banerjee • Spring 2017 • University of Waterloo

Last Revision: January 9, 2018

Table of Contents

Abstract

These notes are intended as a resource for myself; past, present, or future students of this course, and anyone interested in the material. The goal is to provide an end-to-end resource that covers all material discussed in the course displayed in an organized manner. If you spot any errors or would like to contribute, please contact me directly.

1 R Code

1.1 Distribution Commands

There are four functions for most distributions in R. For the Binomial distribution:

`dbinom` Density function. Returns $f(x) = P(X = x) \in [0, 1]$ the probability for a given x value to occur (height of PMF)

$$\mathbb{R}^n \to [0, 1]$$

`pbinom` P-value function (or CDF). Returns p-value or $F(x) = P(X \leq x) \in [0, 1]$ the percentile to which a given x value maps.

$$\mathbb{R}^n \to [0, 1]$$

`qbinom` Quantile function (reverse `pbinom`). Returns the x value that correspond to the p-value or quantile (domain is the range of all values possible for your distribution).

$$[0, 1] \to \mathbb{R}^n$$

`rbinom` Sampling function. Returns n samples from the distribution with the given parameters.

Type `?pbinom` in R console for information on commands.

1.2 θ in exp (Exponential)

In the real world, the parameter $\lambda = \frac{1}{\mu}$ is the rate (where μ is the population mean). Thus if we want 1 sample from the exponential function with $\mu = 5$, we need to call `rexp(1,1/5)`.
In the course, we use $\theta = \frac{1}{\lambda} = \mu$.

2 Data Summaries

2.1 PPDAC

Know the following definitions

Terms
- units - element in a given population (target, study, sample)
- variates - characteristic associated with each unit
- attributes - functions of variates over population

Errors
- sampling - attributes of sample differ from those of study population
- study - attributes of study population differ from those of target population
- measurement - difference in measured and true values

Problem Types
- descriptive - determine attribute of population

- causative - determine existence of causal relationship between two (or more) variates
- predictive - predict response of a variate

Population Types
- target - population we want to describe and produce conclusions for
- study - population available to us in the study
- sample - group of units that we extract variates from

Bias Response bias is the tendency of certain groups of the population to be vocal majorities that may misrepresent the target population

2.2 Summary Techniques

We can summarize a set of data in two ways:

Graphical
- Histogram - replicate density function
- Empirical CDF - compare with theoretical CDF
- Scatter plots - association between two variates
- Box plots - checks distribution. Outliers separate data points $< q(0.25) - 1.5IQR$ or $> q(0.75) + 1.5IQR$.
- Q-Q plots - compare with e.g. normal distribution (linear relationship → Normal)

Numerical
- Central tendency - $\bar{y}$ (sample mean), median, mode
- Variability - s^2, s, range, IQR

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum(y_i - \bar{y})^2 \\ &= \frac{1}{n-1}(\sum y_i^2 - n\bar{y}^2) \end{aligned}$$

- Skewness (mean - median)
- Kurtosis - fatness of tail: higher kurtosis → fatter tails
- Relative Risk - ratio of a given trait between two categories: ≈ 1 means there is no statistical difference, that is independence
- Sample correlation coefficient - $|r| \approx 1$ means high correlation.

$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

3 Point Estimation

3.1 Sample Distribution

$$Y_i \sim f(y_i; \theta)$$

Y_i is the distribution of the ith sample (a sample can be summarized into one number e.g. 2 successes in a Binomial sample of 5 trials results in $y_1 = 2$).
The distribution of these numbers (samples really) follows Y_i.

3.2 Estimate and Estimator

A given parameter θ in n samples varies depending on what our n samples are. The distribution for these θ values is represented as the **point estimator** $\tilde{\theta}$.
A **point estimate** $\hat{\theta}$ is any value that we pick arbitrary from this distribution. Ideally, we want to pick the maximum likelihood estimate (the most probable one based on our samples).

3.3 Likelihood Function

Find the most probable θ that configures our model to have the maximum "chance" of producing our samples.
We combine all our Y_is (combine distributions for each and every sample i) by multiplying their PDFs (**likelihood function**)

$$L(\theta, y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i; \theta)$$

Solve for the maximum value (**maximum likelihood estimate** MLE) by solving for $\frac{dL}{d\theta} = 0$ (first-order condition).
To aid us with exponentials in the PDFs, we can take the log-likelihood or $ln(L)$.

3.4 Invariance Property

If we wanted to find an *attribute of interest* that is a function of unknown parameters, the invariance property states:

Theorem 3.1. If $\hat{\theta}$ is the MLE of θ then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

3.5 Relative Likelihood

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}$$

where $\hat{\theta}$ is the MLE of θ.
The log relative likelihood is $ln(R(\theta))$.

4 Interval Estimation

Two ways to do it: likelihood intervals and "sampling" (coverage and confidence interals).

4.1 Likelihood Interval

For a $100p\%$ LI,

$$\{\theta : R(\theta) \geq p\}$$

We can conclude that values of θ are plausible/implausible based on where they fall in $R(\theta)$ (or if they fall in a certain $100p\%$ LI:

$R(\theta)$	Value of θ is ...
≥ 0.5	very plausible
$[0.1, 0.5)$	plausible
$[0.01, 0.1)$	implausible
< 0.01	very implausible

4.2 Pivotal Quantity and Distribution

We want to map n samples from n Y_i distributions to a **pivotal quantity** that lets us solve for an unknown population parameter. The Central Limit Theorem (CLT - the means of n samples approaches a normal distribution) is very useful.
The known distribution that this pivotal quantity is equivalent to is called the **pivotal distribution**.

4.3 Chi-Squared Distribution

Defined with k degrees of freedom

$$X_k^2 = \sum_{i=1}^{k} Z^2$$

where $Z = G(0,1)$. Note $E(X_k^2) = k$ and $Var(X_k^2) = 2k$.
Note that

$$X_2^2 \sim Exp(2) = \frac{1}{2}e^{-\frac{y}{2}}$$

and for $df > 30$

$$X_n^2 \sim G(n, 2n)$$

where $\sigma^2 = 2n$.
This distribution is used in our pivotal quantity for finding σ for $G(\mu, \sigma^2)$ samples and our LRTS.
The sum of Chi-Squared distributions is Chi-squared. That is

$$X_{k_1}^2 + \ldots + X_{k_n}^2 = X^2_{\sum_{i=1}^{n} k_i}$$

4.4 Exponential and Chi-Squared

Note that for $Y \sim Exp(\theta)$

$$\frac{2Y}{\theta} \sim Exp(2)$$

We know that $X_2^2 \sim Exp(2)$ thus

$$\sum_{i=1}^{n} \frac{2Y_i}{\theta} \sim X_{2n}^2$$

4.5 Student's T-Distribution

This is distribution shows up when we use the sample deviation instead of the population deviation

$$T_k = \frac{Z}{\sqrt{\frac{X_k^2}{k}}}$$

Refer to Pivotal Quantities.

4.6 Coverage/Confidence Interval

Using our pivotal quantity (with the unknown parameter we want to find) and the pivotal distribution, we can bound a coverage (and confidence) interval that the unknown parameter is probable to take on based on our samples. For this example, we n samples taken from $Y_i \sim Bin(n, \theta)$ distributions:

Step 1 We want to construct an interval for unknown parameter θ. Construct our pivotal quantity and distribution

$$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \sim G(0,1)$$

Step 2 For a 100p% coverage interval, we construct the following **two-tailed** interval

$$\begin{aligned}
&P(-z^* \leq Z \leq z^*) = p \\
&P(-z^* \leq \frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \leq z^*) = p \\
&P(\tilde{\theta} - z^*\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}} \leq \theta \leq \tilde{\theta} + z^*\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}) = p
\end{aligned}$$

For a **two-tailed** interval, we want to range in between $-z^*$ and z^* in $Z = G(0,1)$ to contain p proportion of the distribution.

Thus we take the z-scores at $\frac{1-p}{2}$ $(-z^*)$ and $1-\frac{1-p}{2}$ (z^*). For $p = 0.95$, this corresponds to z-scores p-values 0.025 and $p = 0.975$ (that is ± 1.96).

Step 3 To find the $100p\%$ *confidence* interval, we use the MLE $\hat{\theta}$ in place of $\tilde{\theta}$. Thus our CI for θ for $\hat{\theta} = \frac{\bar{y}}{n}$

$$[\hat{\theta} - z^*\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}, \hat{\theta} + z^*\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}]$$

4.7 Sample Size in Binomial Sampling

Sometimes we want to guarantee a range for θ with $100p\%$ confidence with Binomial samples by adjusting the sample size n.
Note the $100p\%$ confidence interval for our binomial samples $Y_i \sim Bin(n,\theta)$ is

$$\hat{\theta} \pm z^*\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

The $\pm$ part defines the length of our interval or the **margin of error** (% of mean). Ideally for an interval length of less than l (on *one side*)

$$\begin{aligned}
&z^*\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq l \\
&n \geq (\frac{z^*}{l})^2 \cdot \frac{1}{4}
\end{aligned}$$

where $\frac{1}{4}$ comes from noting that arbitrary value $\hat{\theta}(1-\hat{\theta})$ takes on a maximum value of $0.5(1-0.5) = 1/4$.

4.8 Number of Samples for Gaussian and Other Distributions

A similar method as above to bound the margin of error can be applied to the CI of Gaussian and other distributions.

Note the one-sided length of a Gaussian CI is

$$z^* \frac{\sigma}{\sqrt{n}}$$

Thus we can upper bound this by l the margin of error and solve for n.

5 Hypothesis Testing

The gist is to make an assumption (**null hypothesis** H_0) about the parameters of n samples and conclude whether it is plausible or not.
For example, for $Y_i \sim Bin(n, \theta)$, we may assume that $P(\text{success}) = 0.5$ thus our hypothesis is

$$H_0 : \theta = 0.5$$
$$H_1 : \theta \neq 0.5$$

where H_1 is our **alternative hypothesis**.
To quantitatively test our hypothesis, we employ a test statistic.

5.1 Test Statistic

Ideally, our test statistic or discrepancy measure D (some distribution) with discrepancy value d is a distribution that:

(i) $D \geq 0$

(ii) $D = 0$ implies best evidence for H_0

(iii) Larger values of D, stronger evidence against H_0

(iv) $P(D \geq d)$ is our p-value and can be calculated assuming H_0 is true

Remember we derived many pivotal quantities for all types of distributions. Since many of these follow a $G(0, 1)$ or T distribution, we must take the absolute distribution.
For example, the D for Gaussian samples may be (where $H_0 : \mu = \mu_0$)

$$D = \left| \frac{\bar{Y} - \mu_0}{\frac{S}{\sqrt{n}}} \right| \sim |G(0, 1)|$$

Note the p-value calculations must undo the absolute sign!

$$P(D \geq d)$$
$$P(|Z| \geq d) \qquad\qquad 2(1 - P(Z \leq d))$$

the last statement follows by taking the tail sides of $P(Z \leq -d)$ and $P(Z \geq d)$.
Note of *one-sided* hypothesis tests, we need only take one tail side.
A hypothesis is plausible/implausible based on the p-value:

p-value	there is ... against H_0
> 0.1	no evidence
$(0.05, 0.1]$	weak evidence
$(0.01, 0.05]$	strong evidence
≤ 0.01	very strong evidence

The p-value can be interpreted as how unusual (smaller the p-value, the more unusual) our evidence/sample is assuming H_0 is true.
Generally we reject H_0 is p-value is ≤ 0.1, but this depends on the context.

5.2 Test Statistic for Variance in Gaussian

Recall we have the pivotal quantity for σ^2 in n Gaussian samples which satisfies all properties of D. For $H_0 : \sigma = \sigma_0$

$$D = \frac{(n-1)S^2}{\sigma_0^2} = X_{n-1}^2$$

so for our discrepancy value we have

$$d = \frac{(n-1)s^2}{\sigma_0^2}$$

When we compute the p-value, note that the Chi-Squared distribution is not symmetric. To take into account large and small values of d that provide evidence against H_0, we multiply the smaller side by two. We have two cases:

$P(X_{n-1}^2 \leq d) < \frac{1}{2}$ (d is "small"): we take $2P(X_{n-1}^2 \leq d)$

$P(X_{n-1}^2 \leq d) > \frac{1}{2}$ (d is "large"): we take $2(1 - P(X_{n-1}^2 \leq d))$ or $2P(X_{n-1}^2 \geq d)$

5.3 Confidence and p-value

The p-value was derived with respect to an interval, which can be mapped to confidence intervals. That is: a parameter value $\theta = \theta_0$ falls in a $100q\%$ confidence interval for θ *if and only if* the p-value for testing $H_0 : \theta = \theta_0$ is greater than or equal to $1 - q$.

5.4 Likelihood Ratio Test Statistic (LRTS)

This test statistic is useful for all types of distributions (assuming n number of samples is large)

$$\Lambda(\theta) = -2ln(R(\theta)) \sim X_{df}^2$$

where df is the degree of freedoms. df is simply the number of unknowns in your distributions (so for $N(\mu, \sigma^2)$ both unknown, we have $df = 2$).
For most cases, we have one unknown paramater θ thus $\Lambda(\theta) \sim X_1^2$.

6 Regression

6.1 Simple Linear Regression Model (SLRM)

We want to model dependent variate Y_i based on explanatory variate x_i (per each ith sample).
If we think of Y_i as a linear function of x_i with intercept α and slope β (both estimators) with a residual (or noise) $R_i \sim G(0, \sigma)$, then we get the following model

$$Y_i = \alpha + \beta x_i + R_i \sim G(\alpha + \beta x_i, \sigma)$$

where $\mu(x_i) = \alpha + \beta x_i$.
The **Gauss-Markov** assumptions state:

(i) Y_i are all independent and normally distributed given x_i (for a given x_i, Y_i is normally distributed)

(ii) $E(Y_i) = \alpha + \beta x_i$ (mean is a linear function of x_i)

(iii) $Var(Y_i) = \sigma^2$ for all i. Variance of each Y_i or residual R_i have the same variance.

The MLEs for the coefficients are:

$$\tilde{\beta} = \frac{S_{xy}}{S_{xx}}$$
$$\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{x}$$
$$\tilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{\alpha} - \tilde{\beta}x_i)^2$$

The unbiased estimator for σ^2 is actually S_e^2 (estimator for standard error)

$$S_e^2 = \frac{1}{n-2}\sum(Y_i - \tilde{\alpha} - \tilde{\beta}x_i)^2 = \frac{1}{n-2}(S_{yy} - \tilde{\beta}S_{xy})$$

6.2 Distribution of $\tilde{\beta}$

Note that we can derive the distribution for $\tilde{\beta}$ by letting $a_i = \frac{(x_i-\bar{x})}{S_{xx}}$ (constant) and $\tilde{\beta} = \sum a_i Y_i$. Thus $E(\tilde{\beta}) = \beta$ and $Var(\tilde{\beta}) = \frac{\sigma^2}{S_{xx}}$.
The pivotal quantity for $\tilde{\beta} = G(\beta, \frac{\sigma}{\sqrt{S_{xx}}})$ is

$$\frac{\tilde{\beta} - \beta}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim G(0, 1)$$

and for the variance

$$\frac{(n-2)S_e^2}{\sigma^2} \sim X_{n-2}^2$$

With S_e

$$\frac{\tilde{\beta} - \beta}{\frac{S_e}{\sqrt{S_{xx}}}} \sim T_{n-2}$$

6.3 Distribution of $\tilde{\mu}(x)$

Note that $\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x_i$ or the sum of Gaussian distributions, thus $\tilde{\mu}(x) = G(\mu(x), \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}})$.
The pivotal quantities are

$$\frac{\tilde{\mu(x)} - \mu(x)}{\sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0, 1)$$
$$\frac{\tilde{\mu(x)} - \mu(x)}{S_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim T_{n-2}$$

where $\mu(x) = \alpha + \beta x$.
For the distribution of $\tilde{\alpha}$, plug in $x = 0$ into the above.

6.4 Prediction Interval for Y_{new}

Note that $Y_{new} \sim G(\alpha + \beta x_{new}, \sigma)$ and $\tilde{\mu}_{new} \sim G(\alpha + \beta x_{new}, \sigma\sqrt{\frac{1}{n} + \frac{(x_{new}-\bar{x})^2}{S_{xx}}})$. Thus

$$Y_{new} - \tilde{\mu}_{new} = G(0, \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_{new}-\bar{x})^2}{S_{xx}}})$$

with pivotal quantity (for which we can solve for Y_{new}

$$\frac{Y_{new} - \tilde{\mu}_{new}}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_{new}-\bar{x})}{S_{xx}}}} \sim G(0,1)$$

or with S_e

$$\frac{Y_{new} - \tilde{\mu}_{new}}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x_{new}-\bar{x})}{S_{xx}}}} \sim T_{n-2}$$

6.5 Graphical Checks of SLRM Assumptions

Note that $\hat{r}_i^* = \frac{\hat{r}_i}{s_e}$ is the standardized residual.

Scatterplot Should follow a linear relationship (x_i, Y_i)

Residual plots For either $(x_i, \hat{r}_i^*)$ or $(\hat{\mu}_i, \hat{r}_i^*)$, the plots of $\hat{r}_i$ should form a narrow band of values between $[-3, 3]$ with no apparent pattern (homoscedascity).

Q-Q plot against Z Plot quantiles of $\hat{r}_i^*$ against that of Z should be linear near the middle.

7 Goodness of Fit

7.1 LRTS for Multinomial

The multinomial LRTS is useful for a lot of hypothesis testing problems that involve multiple categories. For n sample with distribution $X_i \sim f(x_i; \theta)$

$$\Lambda(\theta) = 2\sum_{j=i}^{k} Y_j ln(\frac{Y_j}{E_j}) \sim X^2_{df}$$

where Y_j are the observed frequencies and E_j are the expected frequencies for category j. We bin the results X_is into k categories. Note that we will collapse 1 or more y_js if $y_j < 5$. df is the degrees of freedom (see below).
E_j is is calculated as

$$E_j = n \times p_j$$

where p_j is the probability a given $x = j$ occurs (or $P(X_i = j)$). Note for λ, we use find $e_j = n \times \hat{p}_j$ using $\hat{\theta}$ in $f(x_i; \theta)$.
For categories with intervals, we need to take the integral for $\hat{p}_j$.
Note $\sum_{j=1}^{k} p_j = 1$ for this to work.
For the p-value, we always take

$$P(\Lambda(\theta) \geq \lambda)$$

instead of the two tailed approach.

7.2 Degrees of Freedom

We have two special cases:

Categorical parameters When we have categorical parameters (like $\theta_j = P(j)$ for face $j \in [1, 6]$ in a dice roll), note that $sum\theta_j = 1$, thus there are only 5 free parameters.

Hypothesis parameters In the null hypothesis, we may assume some parameters (e.g. $H_0 : \theta_1 = 0.5$). We must account for these hypothesis parameters from our df. So for our dice example, we have 5 free parameters, we subtract 1 to account for known θ_1 under H_0, thus $df = 5 - 1 = 4$.

More generally for bounded categorical distributions (that is say the sum of the parameters of the categories is known)

$$df = (n - 1) - p$$

where n is the number of categories and p is the number of parameters in the null hypothesis.

7.3 Testing Two Gaussian Population Means

We want to see test if the mean of two populations with distributions $A_i \sim G(\mu_1, \sigma_1)$ and $B_i \sim G(\mu_2, \sigma_2)$ are equal ($\mu_1 = \mu_2$).
We have three cases:

Matched Data Every A_i is paired with its corresponding B_i. Thus we can take $Y_i = A_i - B_i \sim G(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}$ and test $H_0 : \mu_y = 0$

$$D = |\frac{\bar{Y} - 0}{\frac{S}{\sqrt{n}}}| \sim |T_{n-1}|$$

where $S^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$.

Unmatched Data with Common Variance Note $\sigma_1 = \sigma_2$. We use their sample mean distributions $\bar{Y_1} \sim G(\mu_1, \frac{\sigma}{\sqrt{n_1}})$ and $\bar{Y_2} \sim\sim G(\mu_2, \frac{\sigma}{\sqrt{n_2}})$. Thus we have

$$D = |\frac{(\bar{Y_1} - \bar{Y_2}) - 0}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}| \sim T_{n_1+n_2-2}$$

where

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

Unmatched Data with Different Variances We need $n_1, n_2 \geq 30$ large sample sizes. Similar to how the above derivation with a common variance

$$D = |\frac{(\bar{Y_1} - \bar{Y_2}) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}| \sim |Z|$$

Note it is Z and not T since for large sample sizes n $T_n \sim Z$

7.4 Matched vs Unmatched Testing

Generally, we want matched data since

$$Var(\bar{Y_1} - \bar{Y_2} = Var(\bar{Y_1}) + Var(\bar{Y_2}) - 2Cov(\bar{Y_1}, \bar{Y_2})$$

We expect matched data to have positive covariance (we must enforce this) thus the variance is smaller (ideal). Independent samples imply a covariance = 0, which is worse than matched data since it has a larger variance.

7.5 Independence Testing for Two Variates

For a given population, we may test for the independence of two variates A and B such that they have discrete types/values A_i and A_j. We construct a contingency table with frequencies of each occurrence

$A \setminus B$	B_1	$\dots$	B_b	Total
A_1	y_{11}	$\dots$	y_{1b}	r_1
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
A_a	y_{a1}	$\dots$	y_{ab}	r_a
Total	c_1	$\dots$	c_b	n

Note that independence implies that $P(A_i \cap B_j) = P(A_i) \cdot P(B_j) = \frac{r_i}{n} \cdot \frac{c_j}{n}$.
We can then treat this as a hypothesis testing question with $H_0 : \theta_{ij} = \dots$ and use LRTS

$$\Lambda(\theta) = 2\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij} ln(\frac{Y_{ij}}{E_{ij}}) \sim X^2_{(a-1)(b-1)}$$

where $E_{ij} = n \times P(A_i \cap B_j) = \frac{r_i \times c_j}{n}$.

8 Causation

8.1 Causal Effect and Confounding Variables

We say X has a **causal effect** on Y if all other factors that affect Y are held constant, then a change in X sees a change in Y.
A positive correlation between X and Y can mean at least three things: X causes Y, Y causes X, or some other factor Z causes both X and Y.
A **confounding variable** is any other factors or variates that may affect X or Y.

8.2 Blocking and Randomization in Experimental Studies

For experimental studies, we need to control our confounding variables. There are one of two ways:

Blocking We keep the value/level of the confounding variables constant across all samples

Randomization We randomly partition the samples into our desired categories (e.g. Y and non-Y). We strive to distribute confounding variates evenly.

9 Other Reference Equations

9.1 LI $\iff$ CI

From a $100p\%$ LI to a $100q\%$ CI, we take the q that corresponds to

$$P(-\sqrt{-2lnp} \leq Z \leq \sqrt{-2lnp}) = q$$

From a $100p\%$ CI to a $100q\%$ LI, the relative likelihood ratio value is

$$P(R(\theta) \geq e^{-\frac{z^{*2}}{2}})$$

9.2 Percentile

To find the $100p$th percentile in a sample of $y_1, \ldots, y_n$, we take y_m where

$$m = (n+1)p$$

If $m \notin \mathbb{Z}$, then

$$y_m = \frac{y_j + y_{j+1}}{2}$$

where $j < m < j+1$, $j \in \mathbb{Z}$.

9.3 Relative Risk

Between two discrete (binary) variates, how is one type of variate B affected by the types of A?

$A \setminus B$	B	not B
A	y_{11}	y_{12}
not A	y_{21}	y_{22}

The relative risk of B with respect to A vs not A is

$$\frac{\frac{y_{11}}{y_{11}+y_{12}}}{\frac{y_{21}}{y_{21}+y_{22}}}$$

9.4 PMFs/PDFs

Binomial y number of successes in k # of (Bernoulli) trials and $\theta = P(\text{success})$

$$f(y; k, \theta) = \binom{k}{y} \theta^y (1-\theta)^{k-y}$$
$$\mu = k\theta$$
$$\sigma^2 = k\theta(1-\theta)$$

If k large and θ small, then $Bin(k, \theta) \sim Pois(k\theta)$.

Exponential y is the time between events in a Poisson process where θ is the mean (or inverse rate, where rate is the equivalent of average time in between events)

$$f(y; \theta) = \frac{1}{\theta} e^{-\frac{y}{\theta}}$$
$$\mu = \theta$$
$$\sigma^2 = \theta^2$$

Poisson y is the number of events that occur in an interval where θ is equivalent to the average number of times an event occurs in an interval

$$f(y;\theta) = \frac{\theta^y e^{-\theta}}{y!}$$
$$\mu = \theta$$
$$\sigma^2 = \theta$$

Negative Binomial y is the number of successes before r desired number of failures and $\theta = P(\text{success})$

$$f(y;r,\theta) = \binom{y+r-1}{y}\theta^y(1-\theta)^r$$
$$\mu = \frac{\theta r}{1-\theta}$$
$$\sigma^2 = \frac{\theta r}{(1-\theta)^2}$$

Gaussian/Normal y is the desired value with population mean μ and variance σ^2

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Denoted as $N(\mu,\sigma^2) \sim G(\mu,\sigma)$, where $Z = G(0,1)$.

Geometric y number of failures before first success with $\theta = P(\text{success})$

$$f(y;\theta) = (1-\theta)^y\theta$$
$$\mu = \frac{(1-\theta)}{\theta}$$
$$\sigma^2 = \frac{(1-\theta)}{\theta^2}$$

9.5 MLEs

The maximum likelihood estimate for a given parameter θ is denoted as $\hat{\theta}$

Binomial θ is the mean (or $P(\text{success})$) and k is the number of trials

$$\hat{\theta} = \frac{\bar{y}}{k}$$

Exponential θ is the mean (or inverse rate)

$$\hat{\theta} = \bar{y}$$

Poisson θ is the mean (or average number of times an event occurs)

$$\hat{\theta} = \bar{y}$$

Gaussian/Normal μ is the population mean and σ^2 is the population variance

$$\hat{\mu} = \bar{y}$$
$$\hat{\sigma^2} = \frac{1}{n}\sum(y_i - \bar{y})^2$$

Geometric θ is $P(\text{success})$

$$\hat{\theta} = \frac{1}{\bar{y}+1}$$

9.6 Pivotal Quantities

Binomial

$$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \sim G(0,1)$$

Exponential

$$\frac{\bar{Y}-\theta}{\frac{\bar{Y}}{\sqrt{n}}} \sim G(0,1)$$

Poisson

$$\frac{\bar{Y}-\theta}{\sqrt{\frac{\bar{Y}}{n}}} \sim G(0,1)$$

Gaussian

$$\frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim G(0,1)$$
$$\frac{\bar{Y}-\mu}{\frac{S}{\sqrt{n}}} \sim T_{n-1}$$
$$\frac{(n-1)S^2}{\sigma^2} \sim X^2_{n-1}$$