REVIEW SESSION FOR STA 303

DONGYANG DAWN YANG

PHD CANDIDATE, YEAR 4

FEBRUARY 2, 2022

OVERVIEW

- Statistical concepts
- Quiz questions
- Practice version HAS COMMENTARY!

OUTLINE OF "STATISTICS SO FAR"

Topic	Subtopic
Probability Theory	Probability
	Random Variables and Distributions
	Expectations
	Convergence
Statistical Inference	Models, Statistical Inference and Learning
	Parametric Inference
	Hypothesis Testing and P-values
Regression Models	Linear Regression

PROBABILITY

Sample Spaces (Ω) and Events (A)

- Union ($A \cup B$), Intersection ($A \cap B$), Set difference, Set inclusion
- Disjoint/monotone increasing or decreasing sequence

Definition of Probability

- *Axiom 1*: $P(A) \geq 0 \; \forall \; A$
- *Axiom 2*: $P(\Omega) = 1$
- *Axiom 3*: If $A_1, A_2, \ldots$ are disjoint, then $P(\bigcup_{i=1}^{\infty} A_i) = \Sigma_{i=1}^{\infty} P(A_i)$

Probability on Finite Sample Spaces

- Independence: If A and B are independent $\Leftrightarrow P(AB) = P(A)P(B)$
- Conditional Probability: $P(A \mid B) = P(AB)/P(B)$
- Law of Total Probability: $P(B) = \sum_{i=1}^{k} P(B \mid A_i) P(A_i)$
- Bayes' Theorem: $P(A_i \mid B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_i^k P(B|A_i)P(A_i)}$

RANDOM VARIABLES AND DISTRIBUTIONS

Random Variable. X

Distribution Functions and Probability Functions $F_X(x), f_X(x)$

- Cumulative distribution functions (CDF)
- Probability mass functions (discrete RVs)
- Probability density functions (continuous RVs)

RANDOM VARIABLES AND DISTRIBUTIONS

Discrete Random Variables

- Bernoulli
- Binomial
- Poisson
- Geometric
- Negative Binomial

Continuous Random Variables

- Normal/Gaussian
- Exponential
- Gamma
- Beta
- Student's t
- Chi-square

RANDOM VARIABLES AND DISTRIBUTIONS

Bivariate Random Variables

- Marginal Distributions $f_X(x)$
- Conditional Distributions $f_{X|Y}(x \mid y)$
- Independent Random Variables $f_{XY}(xy) = f_X(x)f_Y(y)$

Multivariate Distributions

- Multinomial distribution
- Multivariate normal distribution

Transformation of Random Variables. $f_{X+Y}(z)$

EXPECTATIONS

Definitions and Properties of:

Expectations	$E[X]$
Variance and Covariance	$\mathrm{Var}(X), \quad \mathrm{Cov}(X, Y)$
Conditional Expectation	$E[\, Y \mid X \,]$
Moment Generating Functions	$M_X(t)$

Application to important discrete and continuous RVs

(Binomial, Normal, Exponential, Poisson)

CONVERGENCE

Type of Convergence

- Convergence in Distribution $X_n \xrightarrow{d} X$
- Convergence in Probability $X_n \xrightarrow{p} X$
- Almost Surely/Everywhere Convergence $X_n \xrightarrow{a.e.} X$

Law of Large Numbers

$$\bar{X}_n \xrightarrow{p} E[X] = \mu$$

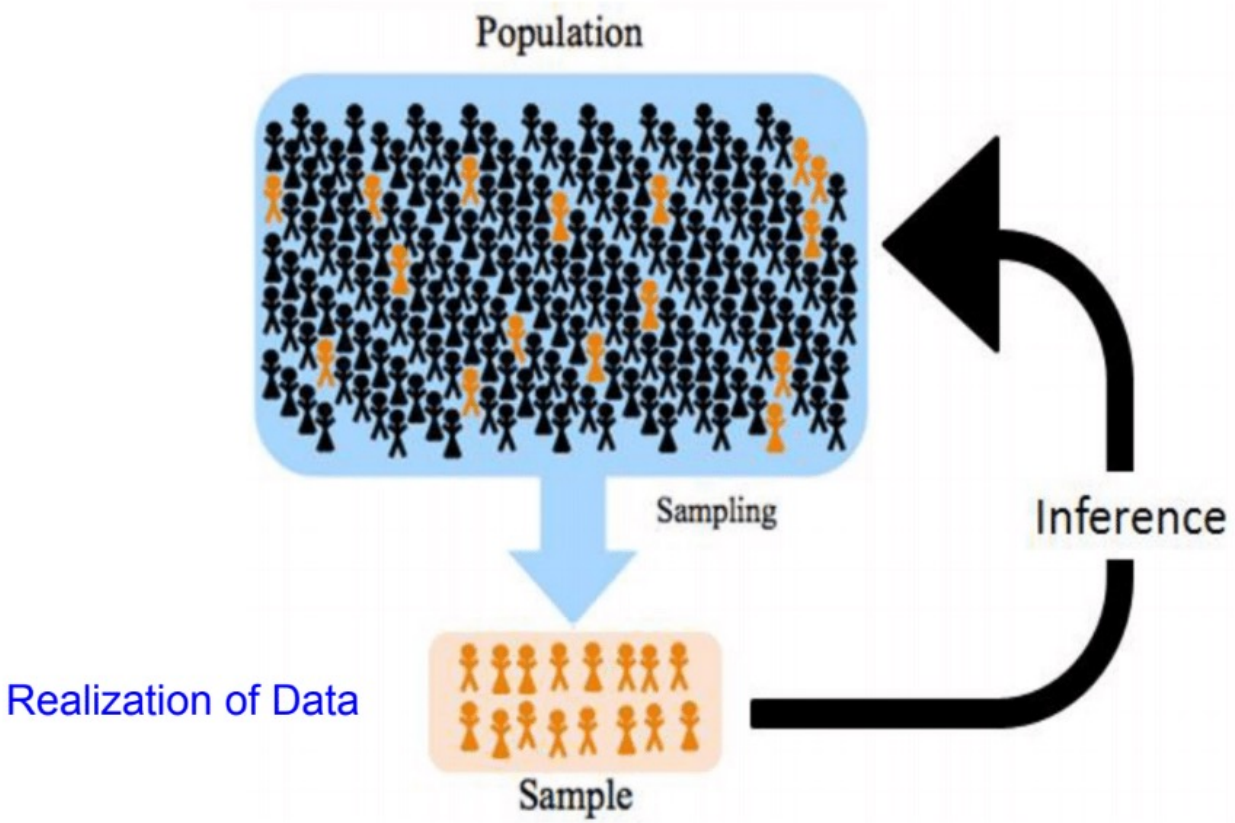
Central Limit Theorem

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim N(0,1)$$

Delta Method

$$\frac{\sqrt{n}\left(g(\bar{X}_n) - g(\mu)\right)}{|g'(\mu)|\sigma} \xrightarrow{d} Z \sim N(0,1)$$

MODELS, STATISTICAL INFERENCE AND LEARNING



Two key questions:

1. What do you infer
2. How do you infer

Source of the figure: https://newonlinecourses.science.psu.edu/stat200/lesson/1/1.2

MODELS, STATISTICAL INFERENCE AND LEARNING

Foundational Concepts in Inference

- Point Estimation $\hat{\theta}$
- Sampling Distribution $\hat{F}$
- Standard Error $\mathrm{se}(\hat{\theta})$
- Confidence Intervals $CI(\hat{\theta}; 0.95)$

"With 95% confidence we can claim that the true value is between (xxx, xxx)."

- Bias, Variance, MSE $\hat{\theta} - \theta,\ \mathrm{Var}(\hat{\theta}),\ [bias(\hat{\theta})]^2 + \mathrm{Var}(\hat{\theta})$
- Consistency $\hat{\theta} \xrightarrow{p} \theta$

REVIEW OF QUIZ

Question 9 1 pts

Which ONE of the following is a correct description of what a 95% confidence interval gives us?

- ○ A range of plausible values for the sample statistic.
- ○ A range of numbers that the population parameter has a 95% chance of being within.
- ○ A range of numbers that the sample statistic has a 95% chance of being within.
- ○ A range of plausible values for the population parameter.

PARAMETRIC INFERENCE

Methods for constructing estimators (and their variances)

Method of Moments

Maximum Likelihood

- The Likelihood Function
- Maximum Likelihood Estimator (MLE)
- Properties of MLE (consistency/equivariance/asymptotic normality/optimality)

The Delta Method

- Deriving Asymptotic Variances
- Single & Multiparameter Models

REVIEW OF QUIZ

Question 11 1 pts

Which of the following claims about the likelihood function are true?

- ◯ The likelihood function is the probability of observing the data we have observed, given a specific model and set of parameters.
- ◯ The statements listed are equivalent.
- ◯ The likelihood function is the probability of a specified model and set of model parameters being appropriate, given the observed data.

HYPOTHESIS TESTING AND P-VALUES

Types of Hypotheses and Errors

- Null hypothesis vs Alternative
- Type I and Type II

Hypothesis Testing

- Difference/Ratio of two means
- Pearson's chi-square test for multinomial data
- The permutation test
- The likelihood ratio test
- Goodness-of-fit tests

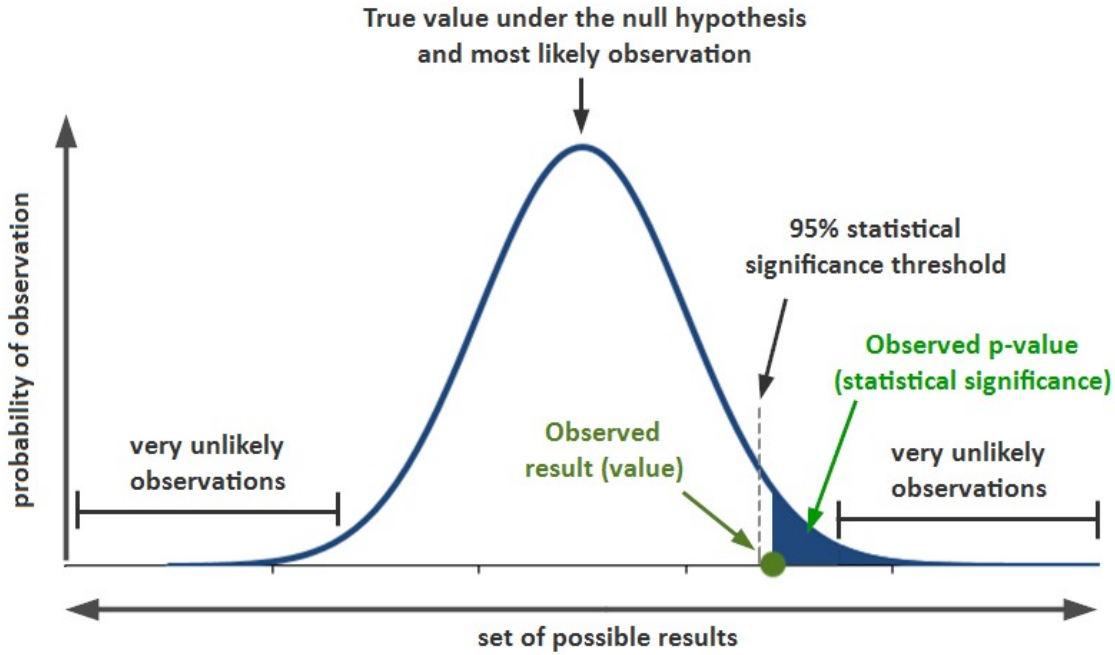
HYPOTHESIS TESTING AND P-VALUES

Hypothesis Testing

- P-values

We should never be making claims in favour/ against the alternative hypothesis.

Our statistical claims are always about the null.



HYPOTHESIS TESTING AND P-VALUES

Hypothesis Testing

- P-values

P-value range	Strength comment
> 0.1	No evidence against the null
0.05 < p-value < 0.1	Weak evidence against the null hypothesis
0.01 < p-value < 0.05	Moderate/some evidence against the null hypothesis
0.001 < p-value < 0.01	Strong evidence against the null hypothesis
< 0.001	Very strong evidence against the null hypothesis

REVIEW OF QUIZ

Input

```
summary(lm(flipper_length_mm ~ bill_length_mm*species, data = penguins))
```

Output

```
Call:
lm(formula = flipper_length_mm ~ bill_length_mm * species, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max
-24.0561  -3.3195  -0.1806   3.5830  16.4397

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  158.9244     6.9039  23.020  < 2e-16 ***
bill_length_mm                 0.7999     0.1776   4.505 9.17e-06 ***
speciesChinstrap             -12.2886    12.4595  -0.986   0.3247
speciesGentoo                 -7.8284    10.6428  -0.736   0.4625
bill_length_mm:speciesChinstrap   0.2073     0.2765   0.750   0.4538
bill_length_mm:speciesGentoo      0.5913     0.2459   2.405   0.0167 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.792 on 336 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8328,    Adjusted R-squared:  0.8303
F-statistic: 334.8 on 5 and 336 DF,  p-value: < 2.2e-16
```

REVIEW OF QUIZ

Question 14 1 pts

Which ONE of the following is the best interpretation of the p-value 0.3247, associated with `speciesChinstrap` in the output above?

- ◯ We have some evidence against the claim that Chinstrap and Adelie penguins have different average flipper lengths.
- ◯ We have evidence in favour of the null hypothesis, that there is no difference in flipper length between Chinstrap and Adelie penguins with the same bill length.
- ◯ We have no evidence against the claim that the coefficient for `speciesChinstrap` is equal to 0.
- ◯ We have strong evidence that we should remove the variable `speciesChinstrap` from the model.

LINEAR REGRESSION

Assumptions for linear regression

- Linear
- Covariance of the errors
- Common error variance
- Normality of errors

Impact of violated assumptions

Supplementary material:

https://bookdown.org/roback/bookdown-BeyondMLR/ch-MLRreview.html

LINEAR REGRESSION

Inference In Regression

- Inference on the Coefficients
  - Continuous
  - Binary/categorical
  - Interactions between continuous and categorical

  Eg. The average increase in arterial blood pressure (Y) for each sqm increase in body surface area (X1) is 1.61 less for males comparing to females (X2), holding other variables in the model constant

REVIEW OF QUIZ

Input

```
summary(lm(flipper_length_mm ~ bill_length_mm*species, data = penguins))
```

Output

```
Call:
lm(formula = flipper_length_mm ~ bill_length_mm * species, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max
-24.0561  -3.3195  -0.1806   3.5830  16.4397

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               158.9244     6.9039  23.020  < 2e-16 ***
bill_length_mm              0.7999     0.1776   4.505 9.17e-06 ***
speciesChinstrap          -12.2886    12.4595  -0.986   0.3247
speciesGentoo              -7.8284    10.6428  -0.736   0.4625
bill_length_mm:speciesChinstrap   0.2073     0.2765   0.750   0.4538
bill_length_mm:speciesGentoo      0.5913     0.2459   2.405   0.0167 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.792 on 336 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8328,    Adjusted R-squared:  0.8303
F-statistic: 334.8 on 5 and 336 DF,  p-value: < 2.2e-16
```

REVIEW OF QUIZ

Question 15 1 pts

Which ONE of the following is an appropriate interpretation of the coefficient for `bill_length_mm:speciesGentoo` in the above output?

- ○ The **slope** for the linear relationship between bill length and flipper length is 0.59 mm per mm steeper for Gentoo penguins than for Adelie penguins.
- ○ The **intercept** for the Gentoo penguins is 0.59 mm higher than for the baseline penguins.
- ○ The **intercept** for the linear relationship between bill length and flipper length is 0.59 mm for Gentoo penguins.
- ○ The **slope** for the linear relationship between bill length and flipper length is 0.59 mm per mm for Gentoo penguins.