Recap of Week 9

- **Wald & Score test**
  - Wald test: MLE & observed Fisher Info → test stat
  - Score test: use distribution of Score function
- **Likelihood ratio test (LRT)**
  - LRT for single population
  - LRT for two populations

    (Note: difference in Λ calculation; df of χ^2 distribution)
- **Goodness of Fit (GOF) test**
  - Intuition of test stat:

$$X^2 = \sum_{i=1}^{k} \frac{(\text{Observed count of } i - \text{Expected count of } i)^2}{\text{Expected count of } i} \xrightarrow{D} \chi^2_{(k-1)}$$

Wald Test Question

Suppose $X_1, X_2, \ldots, X_n$ are independently drawn from an *Exponential*(θ) distribution, where θ is the unknown parameter. Construct a Wald test for testing $H_0: \theta = \theta_0$ vs. $H_1: \theta \neq \theta_0$ at a given α.

(Constructing a test means providing the form of the test statistic, appropriate distribution, and the rejection region.)

Wald Test Question cont'd

Suppose the following observations are independently drawn from the Exponential distribution:

```
> random_values <- rexp(10, rate = 1 / theta)
> random_values <- round(random_values, 2)
> random_values
[1] 3.05 5.61 0.47 0.03 4.19 1.18 5.24 1.77 0.14
2.00
> sum(random_values)
[1] 23.68
```

Test H_0: $\theta = 2$ vs. H_1: $\theta \neq 2$ at $\alpha = 0.05$ using the Wald test that you constructed in the previous question.

LRT Question

Suppose $(X_1, X_2, \ldots, X_n) \overset{\text{iid}}{\sim} Poisson\ (\lambda_x)$ and $(Y_1, Y_2, \ldots, Y_m) \overset{\text{iid}}{\sim} Poisson\ (\lambda_y)$ and they are independent of each other. Construct a likelihood ratio test, for testing the hypothesis, $H_0: \lambda_x = \lambda_y$ vs. $H_a: \lambda_x \neq \lambda_y$ at significance level α. Give the test statistic, rejection region, and distribution.

LRT Question cont'd

Now suppose we generate 3 samples from *Poisson* (λ_x) and 3 samples from *Poisson* (λ_y). The table below shows the results:

X	$1, 5, 0$	$\sum_{i=1}^{3} x_i = 6$	$\bar{x} = 2$
Y	$2, 4, 6$	$\sum_{i=1}^{3} y_i = 12$	$\bar{y} = 4$

Test H_0: $\lambda_x = \lambda_y$ vs. H_a: $\lambda_x \neq \lambda_y$ at $\alpha = 0.05$.

GOF Question

Suppose the following 20 observations are independently drawn from a distribution in R:

0.94, 0.74, 2.00, 0.38, 1.75, 1.84, 0.87, 1.03, 1.01, 0.27, 1.12, 1.90, 1.84, 0.17, 0.15, 0.25, 1.72, 0.07, 1.70, 1.92

By grouping the data into four equal-length intervals, test whether this data is from a *Uniform* [0,2] distribution.

Practice: test whether it is from a *Uniform* [0,θ] distribution.

Recap of Week 10

- **Bayesian Inference**
  - Parameters are RVs following some distributions
- **Prior, Likelihood, and Posterior**
  - **Prior** (***π*(*θ*)**, valid pdf of θ): distribution of the parameter
    - **Types of Prior:**
      1) Conjugate Prior: the prior resulting in a posterior distribution that belongs to the same family of distribution as the prior
      2) Improper Prior: not a valid pdf ($\int_{\Omega} \pi(\theta)\, d\theta \neq 1$)
      3) Non-informative Prior: not adding information to the analysis (the posterior is completely based on the data)

- **Likelihood**: $\boldsymbol{L(s|\theta)} = f(x_1, x_2, \ldots, x_n|\theta)$; s is the data
- **Posterior** ($\boldsymbol{\pi(\theta|s)}$)
  - Calculation → find the **distribution** of posterior

$$\pi(\theta|\boldsymbol{s}) = \frac{L(\boldsymbol{s}|\theta) * \pi(\theta)}{m(\boldsymbol{s})}$$

    - m(s): the marginal distribution of the data s; constant
    - $\pi(\theta|\boldsymbol{s}) \propto L(\boldsymbol{s}|\theta) * \pi(\theta)$ (i.e., Posterior ∝ Likelihood * Prior)
  - **Estimation** using posterior distribution
    - Posterior mean, variance, and median
    - Posterior mode: the value of θ at which the posterior density is the maximum
      - Under non-informative prior, posterior mode ≡ MLE

- **Credible Interval, HPD Interval**
  - Credible interval (or region): Bayesian equivalent idea of confidence interval
  - Highest posterior density (HPD) interval: narrowest

Bayesian Inference Question

Suppose (X_1, X_2, …, X_n) are independently distributed as Poisson(λ) with λ believed to come from a Gamma(α_0, β_0) distribution.

1. Find the posterior distribution of λ.
2. Calculate the posterior mean and variance of λ.
3. Using a non-informative prior, find the posterior mode.

Recap of Week 11

- Correlation & Least Square Regression
  - Correlation Coefficient (r): linear relationship between two variables
    - $-1 \le r \le 1$
      i. r = -1: perfect negative correlation
      ii. r = 0: zero correlation
      iii. r = 1: perfect positive correlation

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

  - Least Square Regression
    - Regression line: $\hat{y} = \beta_2 x + \beta_1$; β_2 - slope, β_1 - y-intercept
    - To estimate the coefficients, use least square method (i.e., minimize the square error $\Sigma(y_i - b_1 - b_2 x_i)^2$)
    - Best estimates:

$$b_1 = \overline{y_i} - b_2\overline{x}$$

$$b_2 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{\sum x_i y_i - n\overline{x}\overline{y}}{\sum x_i^2 - n\overline{x}^2}$$

- Linear regression under Normal distribution
  - Assumption: $(Y \mid X = x) \sim N(\beta_1 + \beta_2 x, \sigma^2)$
  - Use the MLE method with likelihood function

$$L(\beta_1, \beta_2, \sigma^2) = \prod_{i=1}^{n} f(y_i|x_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum(y_i-\beta_1-\beta_2 x_i)^2\right]$$

$$b_1 = \overline{y_i} - b_2\overline{x}$$

$$\Rightarrow \quad b_2 = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2} = \frac{\sum x_i y_i - n\overline{x}\overline{y}}{\sum x_i^2 - n\overline{x}^2}$$

  - Estimators: B_1, B_2
    - B_1, B_2 have Normal distribution

$$B_1 \sim N\left(\beta_1, \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i-\overline{x})^2}\right]\right)$$

$$B_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i-\overline{x})^2}\right)$$

    - Confidence Interval/t-test for β_2

- Coefficient of Determination
  - TSS = RSS + ESS (Total SS = Regression SS + Residual SS)
  - R^2 = RSS/TSS
  - Interpretation: the proportion of variation in Y that can be explained by the model
- Regression with Categorical X variable
  - Simple case: X has only two categories (e.g., male & female, and Y indicates height)
    - Create dummy variables (e.g., $X_f = 1$, if female; $X_f = 0$, if male)
    - Apply regression, interpretation of b_1 and b_2:
      - b_1: average height of a male
      - b_2: average difference of height between male and female
  - General case: multiple comparisons problem
    - X has N categories → need N-1 dummy variables

Correlation & Regression Question

Given the summary of 5 observations below:

$\sum x_i$	$\sum y_i$	$\sum(x_i - \bar{x})^2$	$\sum(y_i - \bar{y})^2$	$\sum(x_i - \bar{x})(y_i - \bar{y})$
15	42	30	32	25

Suppose we are fitting this regression model:

$$(Y \mid X = x) \sim N(\beta_1 + \beta_2 x, \sigma^2)$$

1. Calculate the maximum likelihood estimate of β_1 and β_2, then interpret the meaning of these 2 values.
2. Calculate the value of the coefficient of determination and interpret the value.
3. Construct a 95% CI for β_2
4. Test H_0: $\beta_2 = 0$ vs. H_a: $\beta_2 \neq 0$ for $\alpha = 0.05$. Write a complete conclusion.