Finals Revision Guide

30.003 Probability and Statistics, Term 4 2019

Wei Min Cher

05 Jan 2020

Contents

1 W8: Statistics and Their Distributions

1.1 Definitions

- Population: all observations
- Sample: subset of population
- Random sample: made up of random variables that are independently and identically distributed
- Statistic: quantity whose value can be calculated from sample data
  - A random variable

1.2 Order statistic

For iid RVs $X_1, X_2, \ldots, X_n$ of unknown distribution, they can be rearranged in an increasing order:

$$X_{(1)} \leq X_{(2)} \leq \ldots X_{(k)} \ldots \leq X_{(n)}$$

where

- $X_{(1)} = \min\{X_1, \ldots, X_n\}$ is the smallest order statistic;
- $X_{(k)}$ is the k-th order statistic; and
- $X_{(n)} = \max\{X_1, \ldots, X_n\}$ is the largest order statistic.

1.3 Sample range

The sample range R is the distance between the largest and smallest order statistic.
It is also a random variable, and can be calculated by:

$$R = X_{(n)} - X_{(1)}$$

1.4 Distribution of a statistic

The distribution of a statistic can be obtained by either 1 of the 2 methods:

1. Derive the probability distribution analytically via order statistics
2. Simulate the probability distribution using Monte Carlo simulation

1.5 Distribution of $\overline{X}$

For a sufficiently large n, i.e. $\mathbf{n \leq 30}$, $\overline{X}$ has approximately a normal distribution with mean $E(\overline{X})$ and variance $V(\overline{X})$ as follows:

- Mean, $E(\overline{X}) = \mu$
- Variance, $V(\overline{X}) = \dfrac{\sigma^2}{n}$

1.6 Distribution of smallest order statistic $X_{(1)}$

- pdf:

$$f_{(1)}(x) = n\left[1 - F_X(x)\right]^{n-1} f_X(x)$$

- cdf:

$$\begin{aligned}
F_{(1)}(x) &= P(X_{(1)} \leq x) \\
&= 1 - P(X_i > x, \forall i) \\
&= 1 - \prod_{i=1}^{n} P(X_i > x) \quad \text{(independent)} \\
&= 1 - \left[P(X_i > x)\right]^n \quad \text{(identically distributed)} \\
&= 1 - \left[1 - P(X_i \leq x)\right]^n \\
&= 1 - \left[1 - F_X(x)\right]^n
\end{aligned}$$

1.7 Distribution of largest order statistic $X_{(n)}$

- pdf:

$$f_{(n)}(x) = n\left[F_X(x)\right]^{n-1} f_X(x)$$

- cdf:

$$\begin{aligned}
F_{(n)}(x) &= P(X_{(n)} \leq x) \\
&= P(X_i \leq x, \forall i) \\
&= \prod_{i=1}^{n} P(X_i \leq x) \quad \text{(independent)} \\
&= \left[P(X_i \leq x)\right]^n \quad \text{(identically distributed)} \\
&= \left[F_X(x)\right]^n
\end{aligned}$$

1.8 Distribution of k-th order statistic $X_{(k)}$

- pdf:

$$f_{(k)}(x) = \frac{n!\left[F_X(x)\right]^{k-1}\left[1 - F_X(x)\right]^{n-k} f_X(x)}{(k-1)!(n-k)!}$$

2 W9: Point Estimation

2.1 Point estimate

- Statistic, function of data to infer value of unknown parameter
- A random variable
  - e.g. point estimate of θ is $\hat{\theta}$

2.2 Principle of Unbiased Estimator

- Choose an unbiased estimator among several candidates
- Point estimate $\hat{\theta}$ is an unbiased estimator if $E(\hat{\theta}) = \theta$ for every possible value of θ
- Can be obtained from biased estimator by using making $E(\hat{\theta}) = \theta$

2.3 Principle of Minimum Variance Unbiased Estimation

- Among all the unbiased estimators of θ, choose the estimator with the minimum variance.
- Estimator with the minimum variance is the minimum variance unbiased estimator (MVUE) of θ.

3 W9: Method of Moments Estimator (MME)

3.1 Moments

- k-th population moment, $\mu_k = E(X^k)$
  - Depends on unknown parameters
- k-th sample moment, $M_k = \dfrac{1}{n}\sum_{i=1}^{n} X_i^k$
  - Function of random sample

3.2 Method of Moments

- Assumes that sample moments provide good estimates of the corresponding population moments

3.3 Method of Moments Estimator (MME)

To calculate the MME(s) of θ:

1. Find m population moments, where m is the number of unknown parameters.
2. Find m sample moments.
3. Equate each population moment to its corresponding sample moments
4. Solve for $\theta = (\theta_1, \ldots, \theta_m)$ to obtain the MMEs for θ.

4 W10: Maximum Likelihood Estimator (MLE)

4.1 Likelihood function

Let $X_1, X_2, \ldots, X_n$ have a joint pdf or pmf:

$$L(\theta_1, \ldots, \theta_m) = f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m)$$

The likelihood function is given by

$$L(\theta) = P(X_1 = x, \ldots, X_n = x_n) = \begin{cases} \prod_{i=1}^{n} p(x_i, \theta) & \text{for discrete RVs} \\ \prod_{i=1}^{n} f(x_i, \theta) & \text{for continuous RVs} \end{cases}$$

4.2 Maximizing the likelihood

- The maximum likelihood estimator (MLE) $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are values that maximize the likelihood function such that

$$L(\hat{\theta}_1, \ldots, \hat{\theta}_m) \leq L(\theta_1, \ldots, \theta_m)$$

4.3 Maximum Likelihood Estimator (MLE)

To calculate the MLE of θ:

1. Find the likelihood function $L(\theta)$ based on the distribution.
2. Differentiate $L(\theta)$ with respect to θ, and equate the derivative to 0.
   - The natural logarithm of $L(\theta)$ could simplify calculations.
3. Solve for the MLE of θ.
4. Check if the value is maximum by taking the second derivative of $L(\theta)$.

Notes:

- In some cases, calculus-based techniques are not applicable to maximize likelihood function.
- MLE does not guarantee to produce an unbiased estimator.

5 W10: Confidence Interval

- Quantifies the confidence interval of a point estimate $\hat{\theta}$

$$l(X_1, \ldots, X_n) < \hat{\theta}(X_1, \ldots, X_n) < u(X_1, \ldots, X_n)$$

  - where $l(\ldots)$ is the lower bound and $u(\ldots)$ is the upper bound respectively.

- The interval contains θ with a confidence interval p:

$$P\{\theta \in [l(X_1, \ldots, X_n), u(X_1, \ldots, X_n)]\} = p$$

- The confidence interval p is often set to a high value e.g. 0.95, 0.99 in practice

5.1 Equivalent expressions for Confidence Interval

The following expressions are equivalent in describing a 90% confidence interval (CI) for μ.

$$\begin{aligned} P\left(|\overline{X} - \mu| < \frac{1.65\sigma}{\sqrt{n}}\right) &= 0.90 \\ P\left(\overline{X} - \frac{1.65\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{1.65\sigma}{\sqrt{n}}\right) &= 0.90 \\ P\left[\mu \in \left(\overline{X} - \frac{1.65\sigma}{\sqrt{n}}, \overline{X} + \frac{1.65\sigma}{\sqrt{n}}\right)\right] &= 0.90 \end{aligned}$$

- Replace 1.65 with:
  - 1.96 if CI is 95%
    - Closest Z-score of area 0.97500 in standard normal table
  - 2.58 if CI is 99%
    - Closest Z-score of area 0.99500 in standard normal table
  - **Rule of thumb:**
    - Search for Z score of area $p + \frac{1-p}{2}$ in the standard normal table, where p is the CI.

5.2 Interpretation of Confidence Interval

- e.g. 95% CI for μ
  - As the number of samples collected tend to infinity, 95% of the samples will contain μ.

5.3 Properties of Confidence Interval

- As population variance σ increases, the width of CI increases.
- As sample size n increases, the width of CI decreases.
- As the confidence interval p increases, the width of CI increases.
- At a fixed confidence interval,
  - Large width of CI $\rightarrow$ low precision
  - Small width of CI $\rightarrow$ high precision

6 W11: Hypothesis Testing 1

6.1 Statistical hypothesis

- A claim about values of parameters/form of probability distribution

6.2 Null and Alternative Hypotheses

- Null hypothesis, H_0
  - Claim that is initially assumed to be true
  - H_0 is **always** $H_0 : \theta = \theta_0$
- Alternative hypothesis, H_a
  - Claim that contradicts the null hypothesis H_0
  - H_a has 3 forms with implicit hypothesis
    - $H_a : \theta > \theta_0$ (implicit hypothesis: $\theta \leq \theta_0$)
    - $H_a : \theta < \theta_0$ (implicit hypothesis: $\theta \leq \theta_0$)
    - $H_a : \theta \neq \theta_0$ (implicit hypothesis: $\theta = \theta_0$)

6.3 Hypothesis Testing

- Method to decide whether to accept or reject the null hypothesis, H_0
- Comprises 2 components:
  - Test statistic
    - Function of sample data to make a decision
  - Rejection region
    - Set of values for which the null hypothesis H_0 will be rejected
    - If test statistic falls in rejection region, H_0 will be rejected

6.4 Errors in Hypothesis Testing

- Type I error (α): Rejecting the null hypothesis H_0 when H_0 is true

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

- Type II error (β): Accepting the null hypothesis H_0 when H_a is true

$$\beta = P(\text{accept } H_0 \mid H_a \text{ is true})$$

- Good rejection region yields small α and β
  - Typical approach: specify largest value of α that can be tolerated, then back-calculate for the rejection region

6.5 Hypothesis Testing using Rejection Region

1. Figure out appropriate H_0 and H_a.

2. Figure out appropriate test statistic.

$$\overline{X} = \frac{1}{n}\sum X_i \quad \Longrightarrow \quad Z = \begin{cases} \dfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ known} \\ \dfrac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ unknown} \end{cases}$$

3. Calculate the rejection region based on type I error/significance level α:

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

4. Calculate the normalized sample mean z using sample mean $\overline{x}$.

$$z = \begin{cases} \dfrac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ known} \\ \dfrac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ unknown} \end{cases}$$

5. Compare the normalized sample mean z with the rejection region.

   Reject H_0 if z falls in the rejection region.

   - $H_a : \mu < \mu_0$ (lower-tailed test)
     - Rejection region: $Z < -z_\alpha$
   - $H_a : \mu > \mu_0$ (upper-tailed test)
     - Rejection region: $Z > -z_\alpha$
   - $H_a : \mu \neq \mu_0$ (two-tailed test)
     - Rejection region: $Z < -z_{\alpha/2} \ \cup \ Z > z_{\alpha/2}$

7 W11: Hypothesis Testing 2

7.1 Hypothesis Testing of Difference between 2 Populations

1. Figure out appropriate H_0 and H_a.

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 \neq 0$$

2. Figure out appropriate test statistic.

$$\overline{X_1} - \overline{X_2} = \frac{1}{n}\sum(X_{1i} - X_{2i})$$

$$\implies \quad Z = \begin{cases} \dfrac{\overline{X_1} - \overline{X_2}}{\frac{\sigma}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ known} \\ \dfrac{\overline{X_1} - \overline{X_2}}{\frac{s}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ unknown} \end{cases}$$

3. Calculate the rejection region based on type I error/significance level α:

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

4. Calculate the normalized sample mean z using sample mean $\overline{x_1} - \overline{x_2}$.

$$z = \begin{cases} \dfrac{\overline{x_1} - \overline{x_2}}{\frac{\sigma}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ known} \\ \dfrac{\overline{x_1} - \overline{x_2}}{\frac{s}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ unknown} \end{cases}$$

5. Compare the normalized sample mean z with the rejection region.
   Reject H_0 if z falls in the rejection region.

   - $H_a : \mu < \mu_0$ (lower-tailed test)
     - Rejection region: $Z < -z_\alpha$
   - $H_a : \mu > \mu_0$ (upper-tailed test)
     - Rejection region: $Z > -z_\alpha$
   - $H_a : \mu \neq \mu_0$ (two-tailed test)
     - Rejection region: $Z < -z_{\alpha/2} \,\cup\, Z > z_{\alpha/2}$

7.2 P-value

- A probability, calculated assuming that H_0 is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample.

7.3 Hypothesis Testing using P-value

1. Figure out appropriate H_0 and H_a.

2. Calculate the test statistic value of sample z.

$$z = \begin{cases} \dfrac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ known} \\ \dfrac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} & \text{population standard deviation } \sigma \text{ unknown} \end{cases}$$

3. Determine range of test statistic values as contradictory to H_0 as the above value of z.

   - $H_a : \mu < \mu_0$ (lower-tailed test)
     - Range: $Z < z$
   - $H_a : \mu > \mu_0$ (upper-tailed test)
     - Range: $Z > z$
   - $H_a : \mu \neq \mu_0$ (two-tailed test)
     - Range: $Z > z \cup Z < -z$

4. Calculate probability of getting that range, assuming H_0 is true:

   - $H_a : \mu < \mu_0$ (lower-tailed test)
     - P-value $= P(Z < z \mid H_0 \text{ is true})$
   - $H_a : \mu > \mu_0$ (upper-tailed test)
     - P-value $= P(Z > z \mid H_0 \text{ is true})$
   - $H_a : \mu \neq \mu_0$ (two-tailed test)
     - P-value $= P(Z > z \cup Z < -z \mid H_0 \text{ is true})$

5. Compare the P-value against the significance level α.

   - Reject H_0: P-value $\leq \alpha$
   - Accept H_0: P-value $> \alpha$

7.4 Comparison between Hypothesis Testing Methods

- The two procedures – the rejection region method and P-value method – are equivalent.
  - The same conclusion will be reached via either of the two procedures.

8 W12: Linear Regression

8.1 Least-squares method

- Estimates unknown parameters of a function based on known data

8.2 Estimating β_0 and β_1

1. Define an error function to minimize.

$$f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)^2$$

2. Take the partial derivative of the error function with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and solve for the unknowns.

$$\begin{aligned}
\frac{\partial f}{\partial \hat{\beta}_1} = 0 :& -2 \sum (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)(-x_i) = 0 \\
&\sum x_i (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0) = 0 \\
&\Rightarrow \sum (\hat{\beta}_1 x_i^2 + \hat{\beta}_0 x_i) = \sum (x_i y_i)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial f}{\partial \hat{\beta}_0} = 0 :& -2 \sum (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)(-1) = 0 \\
&\sum (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0) = 0 \\
&\Rightarrow \sum (\hat{\beta}_1 x_i + \hat{\beta}_0) = \sum y_i
\end{aligned}$$

$$\text{Design matrix of error function: } \sum_{i=1}^{n} \begin{bmatrix} x_i^2 & x_i \\ x_i & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_0 \end{bmatrix} = \sum_{i=1}^{n} \begin{bmatrix} x_i y_i \\ y_i \end{bmatrix}$$

3. Examine the Hessian matrix to determine if the solutions are at a minimum, i.e.

$$\begin{bmatrix} \dfrac{\partial f}{\partial \hat{\beta}_0^2} & \dfrac{\partial^2 f}{\partial \hat{\beta}_0 \hat{\beta}_1} \\ \dfrac{\partial^2 f}{\partial \hat{\beta}_0 \hat{\beta}_1} & \dfrac{\partial f}{\partial \hat{\beta}_1^2} \end{bmatrix} \text{ is positive definite.}$$

8.3 Least-squares estimates for β_0 and β_1

$$\begin{aligned}
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \\
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\overline{x}^2}
\end{aligned}$$

- $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is called the estimated regression line or least-squares line

8.4 Residuals and fitted values

- Residual, $y_i - \hat{y}_i$
  - The difference between the observed value y_i and the fitted value $\hat{y}_i$
  - Positive residual $\rightarrow$ observed point lies above the least-squares line
  - Negative residual $\rightarrow$ observed point lies below the least-squares line
- Sum of residuals, $y_i - \hat{y}_i$
  - For an estimated regression line obtained by the least-squares method, the sum of residuals is zero:
$$\sum_{i=1}^{n} y_i - \hat{y}_i = 0$$
- Fitted values $\hat{y}_i$
  - Obtained by substituting x_i into the regression line equation:
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

8.5 The simple linear regression model

- The simple linear regression model can be described by the model equation
$$Y = \beta_0 + \beta_1 x + \varepsilon$$
  where ε represents uncertainty of the model and is a normal N(0, σ^2) RV.
- The line $y = \beta_0 + \beta_1 x$ is called the true/population regression line.
- Mean of Y, $E(Y)$
$$\begin{aligned} E(Y) &= E(\beta_0 + \beta_1 x + \varepsilon) \\ &= \beta_0 + \beta_1 x + E(\varepsilon) \\ &= \beta_0 + \beta_1 x \end{aligned}$$
- Variance of Y, $V(Y)$
$$\begin{aligned} V(Y) &= V(\beta_0 + \beta_1 x + \varepsilon) \\ &= 0 + V(\varepsilon) \\ &= \sigma^2 \end{aligned}$$

8.6 Sum of squared error (SSE)

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

- Measures discrepancy between the data and the estimation model
- Small SSE $\rightarrow$ tight fit of estimation model to data

8.7 Estimating σ^2 of regression model

- An unbiased estimate for σ^2 in the regression model is s^2:

$$s^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

- Estimating β_0 and β_1 results in the loss of 2 degrees of freedom
  - Thus the denominator for s^2 is $n-2$