Probability and Statistics Notes

Chapter Six

Jesse Crawford

Department of Mathematics
Tarleton State University

Spring 2011

Outline

Statistical Models

Definition (Informal)

- A *statistical model* is a mathematical framework
- used to model random variables,
- where the probability distribution of the variables is not completely known.
- Often, the random variables represent a *random sample*
- from some *population*, where
  - the parametric form of the population distribution is known, but
  - the actual values of the *parameters* are not.

Important Components of a Statistical Model

- Random variables $X_1, \ldots, X_n$, which represent a *sample* from some population.
- A p.d.f./p.m.f. $f(x;\theta)$, representing the *population* distribution.
- The *unknown parameter* θ, a number related to the population distribution whose value is not known.
- The *parameter space*, Ω, consisting of all possible values of θ.

Example

- In a large city, the proportion of voters who approve of the mayor is unknown.
- A random sample $X_1, \ldots, X_n$ is taken from this city, where
- $X_i = 1$ if the ith voter approves, and $X_i = 0$ otherwise.

1. What p.d.f./p.m.f. should be used to model the population?
2. What type of distribution is this?
3. What is the unknown parameter, and what does it represent?
4. What is the parameter space?
5. Suppose 52% of the sample approves of the mayor. What would be the best estimate for p?

The Likelihood Function

- We can compactly summarize the assumptions of our statistical model as
  - $X_1, \ldots, X_n$ are IID,
  - with *common distribution* $f(x; \theta)$, where $\theta \in \Omega$.
- Therefore, the joint p.d.f./p.m.f. of $X_1, \ldots, X_n$ is

$$L(\theta) = L(\theta, x_1, \ldots, x_n) = f(x_1; \theta) \cdots f(x_n; \theta).$$

- This function is called the *likelihood function*.
- The *log-likelihood* is

$$l(\theta) = \ln[L(\theta)].$$

- Intuitively, $L(\theta, x_1, \ldots, x_n)$ is the likelihood of observing $X_1 = x_1, \ldots, X_n = x_n$ when the true value of the parameter is θ.

Maximum Likelihood Estimation

- Consider a random sample $X_1, \ldots, X_n$ from a statistical model with parameter θ.
- An *estimator* for θ is any function

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

  intended to estimate θ based on the sample observations $X_1, \ldots, X_n$.
- $\hat{\theta}$ is the *maximum likelihood estimator* for θ if

$$L[\hat{\theta}(x_1, \ldots, x_n), x_1, \ldots, x_n] = \max_{\theta \in \Omega} L(\theta, x_1, \ldots, x_n),$$

  for all $x_1, \ldots, x_n \in \mathbb{R}$.
- In other words, for any observations $x_1, \ldots, x_n$, the MLE $\hat{\theta}(x_1, \ldots, x_n)$ is the value of θ that would have given the maximum chance of observing those particular sample values, $x_1, \ldots, x_n$.

Proposition

If the population is Bernoulli(p), then the MLE is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

- Useful when studying a binary characteristic of the population (approves/disapproves of mayor).
- $p =$ the proportion of the population having the characteristic (population proportion).
- $\hat{p} =$ the proportion of the sample having the characteristic (sample proportion).

Example

- A company receives phone calls according to a Poisson process.
- Let $X_1, \ldots, X_n$ be n waiting times between successive phone calls.

1. What p.d.f./p.m.f. should be used to model the population?
2. What type of distribution is this?
3. What is the unknown parameter, and what does it represent?
4. What is the parameter space?
5. What is the MLE for θ?

Example

- In a laboratory, mice have weights that are normally distributed.
- Let $X_1, \ldots, X_n$ be a random sample of n mice.

1. What p.d.f./p.m.f. should be used to model the population?
2. What type of distribution is this?
3. What is the unknown parameter, and what does it represent?
4. What is the parameter space?
5. What is the MLE for (μ, σ^2)?

Unbiasedness

Definition

- Let $\hat{\theta}$ be an estimator for a parameter θ.
- If $E(\hat{\theta}) = \theta$, then $\hat{\theta}$ is called *unbiased*. Otherwise, it is *biased*.

Multiple Regression

- A multiple linear regression model is

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \ i = 1, \ldots, n$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$Y = X\beta + \epsilon$$

- It is assumed that $\epsilon \sim \mathrm{N}(0, \sigma^2 I_n)$
- Y is the observable random vector.
- X can be regarded as an observable constant matrix.
- $\beta \in \mathbb{R}^p$ is an unknown parameter vector.
- The MLE for β is

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

Method of Moments

- Another method for estimating parameters is the *method of moments*.
- Suppose the model has r parameters $\theta_1, \ldots, \theta_r$.
- Equate the first r moments of the distribution to the first r moments of the sample, and solve for the parameters to find estimates for them.

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E(X^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\vdots$$

$$E(X^r) = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

Outline

Normal Population with Known Variance

Confidence Interval for μ when

- Population is N(μ, σ^2), and
- σ is known.

Example

- Consider math SAT scores at a university, and assume
- they are normally distributed,
- the mean is unknown, and
- the standard deviation is known to be 100.
- A random sample of size 200 is taken, and
- the sample mean is 517.
- Estimate the average math SAT score at this university.
- Find a 95% confidence interval for the average math SAT score at the university.

Normal Population with Known Variance

Proposition

- If the population is N(μ, σ^2), and
- the population variance σ^2 is known, then

$$P\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha.$$

- A $1 - \alpha$ confidence interval for μ is

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right], \text{ or equivalently,}$$

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

- The parameter μ is fixed.
- The confidence interval is random.

Normal Population with Unknown Variance

Confidence Interval for μ when

- Population is N(μ, σ^2), and
- σ is unknown.

Example

- Consider verbal SAT scores at a university, and assume
- they are normally distributed,
- with unknown mean and standard deviation.
- A random sample of size 25 is taken,
- the sample mean is 561,
- and the sample standard deviation is 124.
- Estimate the average verbal SAT score at this university.
- Find a 95% confidence interval for the average verbal SAT score at the university.

Normal Population with Unknown Variance

Proposition

- If the population is N(μ, σ^2), and
- the population variance σ^2 is unknown, then

$$P\left[\overline{X} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right] = 1 - \alpha.$$

- A $1 - \alpha$ confidence interval for μ is

$$\left[\overline{X} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right],$$

or equivalently,

$$\overline{X} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}.$$

Large Sample Sizes

Recall that for values of $n > 31$,

$$t(n-1) \approx \text{N}(0,1), \text{ so}$$

$$t_{\alpha/2}(n-1) \approx z_{\alpha/2}.$$

Example

- Gas mileages of a certain type of vehicle are normally distributed.
- Gas mileage measurements are made on 100 vehicles, resulting in
- $\overline{X} = 33.5$ and $s = 5.68$.
- Find a 90% confidence interval for the average gas mileage of all such vehicles.

Approximations for Non-normal Populations

Example

- A sample of 200 mice were exposed to a stimulus,
- and response times were measured, resulting in
- a mean response time of 1.12 seconds,
- and a standard deviation of 0.53 seconds.
- Find a 99% confidence interval for the mean response time in the population.

Approximations for Non-normal Populations

Proposition

- For non-normal populations,
- an approximate $1 - \alpha$ confidence interval for μ is

$$\overline{X} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}},$$

- assuming that one of the following conditions holds:
  - $n \geq 30$, or
  - the population distribution does not depart too far from normality (for example, the approximation should be good for a symmetric, unimodal, continuous population distribution).

One-sided Confidence Intervals

- Interested in a *lower bound* for μ.
- Use the one-sided $1 - \alpha$ confidence interval

$$\left[\overline{X} - t_\alpha(n-1)\frac{s}{\sqrt{n}}, \infty\right).$$

- (Assuming appropriate conditions are met. Use z_α instead where appropriate.)

Example

- Pipes manufactured by a company must have a mean strength $\geq$ 2400 lb/ft.
- In a sample of 150 pipes,
- the mean strength was 2437 lb/ft,
- and the standard deviation was 129 lb/ft.
- Find the relevant one-sided 99% confidence interval for the mean pipe strength.
- Does it appear that the pipes in the population exceed the strength requirement?

One-sided Confidence Intervals

- Interested in an *upper bound* for μ.
- Use the one-sided $1 - \alpha$ confidence interval

$$\left(-\infty, \overline{X} + t_\alpha (n-1) \frac{s}{\sqrt{n}} \right].$$

- (Assuming appropriate conditions are met. Use z_α instead where appropriate.)

Example

- Mean emissions from car engines are required to be $\leq$ 20 ppm of carbon.
- The emissions statistics for a sample of 20 engines were
- $\overline{x} = 19.78$ and $s = 1.84$.
- Find the relevant one-sided 99% confidence interval for the emissions levels.
- Does it appear that the engines in the population meet the emissions standards?

Outline

Normal Populations with Known Variances

- Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are **independent** samples
- from two **normal distributions** $\mathrm{N}(\mu_X, \sigma_X^2)$ and $\mathrm{N}(\mu_Y, \sigma_Y^2)$,
- where the variances σ_X^2 and σ_Y^2 are **known**.
- A $1 - \alpha$ confidence interval for $\mu_X - \mu_Y$ is

$$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\sigma_X^2/n + \sigma_Y^2/m}.$$

Normal Populations with Common Unknown Variance

- Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are **independent** samples
- from two **normal distributions** $\text{N}(\mu_X, \sigma^2)$ and $\text{N}(\mu_Y, \sigma^2)$,
- with **common, unknown** variance σ^2.
- A $1 - \alpha$ confidence interval for $\mu_X - \mu_Y$ is

$$\overline{X} - \overline{Y} \pm t_{\alpha/2}(n + m - 2) S_p \sqrt{\frac{1}{n} + \frac{1}{m}},$$

where S_p is the *pooled estimator* of σ,

$$S_p = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n + m - 2}}.$$

Large Samples

- Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are **large independent** samples ($n, m \geq 30$)
- from two distributions with means μ_X and μ_Y.
- A $1 - \alpha$ confidence interval for $\mu_X - \mu_Y$ is

$$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{S_X^2/n + S_Y^2/m}.$$

- We are **not assuming normality, common variance, or known variance**

Relevance to Studies

- The most reliable type of study is a *randomized controlled experiment*.
- *Controlled* means that at least two groups of subjects are studied, often called a treatment group and a control group.
- An *experiment* is a study where the investigator determines which subjects are in which groups, as opposed to an *observational study*, where the investigator simply observes without intervening.
- An experiment is *randomized* if the investigator assigns subjects to treatment/control groups randomly.
- Medical studies should be *double blind*. Neither the patient nor the doctors measuring responses to treatments should know who received the treatment.
- This requires patients to take *placebos* and separate doctors to administer treatments and measure responses.
- If a group is divided into treatment/control randomly, the resulting samples are not independent, but *they may be treated as such*, because this results in *conservative confidence intervals*.

Paired Observations

- Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are n pairs of measurements
- where $E(X_i) = \mu_X$ and $E(Y_i) = \mu_Y$, for $i = 1, \ldots, n$.
- Let $D_i = X_i - Y_i$, for $i = 1, \ldots, n$.
- Assuming the **populations are normally distributed or the sample size is large** ($n \geq 30$),
- a $1 - \alpha$ confidence interval for $\mu_X - \mu_Y$ is

$$\overline{D} \pm t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}.$$

Outline

CI for Variance of a Normal Population

- Suppose $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$,
- and let a and b be constants such that

$$P[a \leq \chi^2(n-1) \leq b] = 1 - \alpha,$$

  i.e., $a = \chi^2_{1-\alpha/2}(n-1)$ and $b = \chi^2_{\alpha/2}(n-1)$.
- Then a $1 - \alpha$ confidence interval for σ^2 is

$$\left[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a} \right].$$

Outline

Mathematical Framework for Proportions

- Consider a population whose subjects have some binary characteristic (approves of mayor or doesn't).
- The proportion of the population with the characteristic is p, the *population proportion*.
- Mathematically, the population is just Bernoulli(p).
- Let $X_1, \ldots, X_n$ be a sample from this population, and let $Y = \sum_{i=1}^{n} X_i$.
- Then $Y \sim$ Binomial(n, p).
- The MLE for the population proportion p is the sample proportion

$$\hat{p} = \frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}.$$

- Also note that the population mean and variance are

$$\mu = p \text{ and } \sigma^2 = p(1-p).$$

- In particular, a good estimator for σ^2 is

$$\hat{\sigma}^2 = \hat{p}(1-\hat{p}) \approx S^2.$$

- Therefore, **as long as the CLT applies (**$np \geq 5$ **and** $n(1-p) \geq 5$**)**, all inferences for a population proportion are the same as those for a population mean, using the following dictionary:

Means	Proportions
μ	p
$\overline{X}$	$\hat{p}$
S	$\sqrt{\hat{p}(1-\hat{p})}$

Confidence Intervals for Populations Proportions

If $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$, a $1-\alpha$ confidence interval for a population proportion is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

If either $n\hat{p}$ or $n(1-\hat{p})$ is less than 5, replace $\hat{p}$ with

$$\tilde{p} = \frac{Y+2}{n+4}.$$

Difference in Population Proportions

- Consider **independent** samples of sizes n_1 and n_2
- from two populations with proportions p_1 and p_2, respectively.
- A $1 - \alpha$ confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Outline

Sample Size for Population Mean

Example

- Suppose you want to estimate the gas mileage of a certain type of car.
- You want a 95% confidence level that is within 2 mpg of the true gas mileage.
- Based on a preliminary study, the standard deviation of the gas mileages is about 5.68 mpg.
- What sample size is required to obtain the desired confidence interval?

- Letting ε denote the desired *margin of error*, we have

$$\varepsilon = z_{\alpha/2}\frac{s}{\sqrt{n}},$$

- so the necessary sample size is

$$n = \frac{z_{\alpha/2}^2 s^2}{\varepsilon^2}.$$

Sample Size for Population Proportion

Example

- Suppose the unemployment rate has been near 8% recently.
- We wish to estimate the unemployment rate within 0.001 with a 99% confidence level.
- What sample size is required?

$$n = \frac{z^2_{\alpha/2}\hat{p}(1-\hat{p})}{\varepsilon^2}.$$

Population Proportion with No Preliminary Estimate

Example

- Politician is considering running for governor.
- Wants to estimate her approval rating within 0.03 with 95% confidence.
- What sample size is required?

$$n = \frac{z^2_{\alpha/2}}{4\varepsilon^2}.$$

Finite Population Correction Factor

- All of our results so far have assumed an **infinite population**.
- Generally, if the sample size is $\leq 5\%$ of the population size, the population can be regarded as infinite.
- For finite populations, the variance of the estimators $\overline{X}$ and $\hat{p}$ is multiplied by the **finite population correction factor**,

$$\frac{N-n}{N-1},$$

where $N =$ population size, and $n =$ sample size.

Confidence Intervals for Finite Populations

$$\overline{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1} \right)}.$$

Example

- Consider a population of 750 college algebra students.
- Suppose we want to estimate the proportion p of these students who met certain performance standards on their final exams.
- We would like to estimate p within 0.05 with 95% confidence.
- What sample size is required?