Probability and Statistics Notes

Chapter Seven

Jesse Crawford

Department of Mathematics
Tarleton State University

Spring 2011

Outline

- *Statistical testing problems* usually involve two *hypotheses* about a parameter, such as

$$\mathrm{H}_0 : \mu \leq 150 \text{ vs. } \mathrm{H}_1 : \mu > 150.$$

- Here, the parameter μ represents the average mass of an apple produced using a new type of fertilizer.
- Assuming that 150 grams is the mass of an apple produced using the old fertilizer, H_0 represents a hypothesis of no change, and H_0 is called the *null hypothesis*.
- H_1 is the hypothesis that the fertilizer manufacturer would need to prove to convince people to buy this type of fertilizer. H_1 is called the *alternative hypothesis*.

- For mathematical reasons, the testing problem is usually restated as

$$\mathrm{H}_0 : \mu = 150 \text{ vs. } \mathrm{H}_1 : \mu > 150.$$

  (Replaced $\leq$ with $=$ in H_0)
- Two types of errors:
  - Type I: Reject H_0 when H_0 is true.
  - Type II: Don't reject H_0 when H_0 is false.
  - (We never accept H_0)
- We can't decrease the chance of both types of errors without raising the sample size.

- Solution: settle for making P[Type I error] small.
- We choose a number α, called the *significance level* of the test, and construct the test so that

$$P[\text{Type I error}] = \alpha.$$

- A conventional value of α is 0.05, but other values may be more appropriate, depending on the situation.
- If α is small, rejecting H_0 is strong evidence that H_0 is false, and the smaller α is, the stronger the evidence is.
- Unfortunately, not rejecting H_0 is **not** strong evidence that H_0 is true. This is a more delicate matter that's addressed in chapter 10.
- Hypothesis tests with significance level α correspond to confidence intervals with confidence level $1 - \alpha$.

- The remaining slides in this section are not essential to material in chapter 7.
- $$P[\text{Type II error}] = \beta.$$
- Among all tests with significance level α, we would like the one that minimizes β.
- $K = 1 - \beta$ is called the *power* of the test.
- Minimizing β is equivalent to maximizing power, so we want a *most powerful test*.
- Notice that β is actually a function of the unknown parameter.

- For example, let's return to

$$\mathrm{H}_0 : \mu = 150 \text{ vs. } \mathrm{H}_1 : \mu > 150,$$

  and let's assume the population standard deviation is $\sigma = 10$.
- If μ is actually equal to 150.00001, then H_0 is false, but the chance of rejecting H_0 will be low, unless the sample size is extremely large. The chance of making a type II error would be large and the power would be low.
- If μ is actually equal to 300, rejecting H_0 would be very likely, even with a small sample size. The chance of making a type II error would be small and the power would be high.
- This shows how β and the power depend on the unknown parameter.
- The tests studied in this chapter maximize power for each value of the unknown parameter, so they are called *uniformly most powerful tests*, a topic discussed in more detail in chapter 10.

- The power of these tests is still low if the true parameter is close to H_0 (relative to the population variance and the sample size). In these cases, achieving a high power level is not possible.
- (Maximum power does not mean high power. If a high power is impossible, the maximum will still be low.)
- In other words, even if we don't reject H_0, it is always possible that H_0 is false, but the true parameter is so close to H_0 that it appears that H_0 is true.
- For this reason we can never accept H_0.

Outline

- Let μ be the mean of a population, and consider the hypothesis test

$$\mathrm{H}_0 : \mu = \mu_0 \text{ vs. } \mathrm{H}_1 : \mu \neq \mu_0.$$

- Suppose $X_1, \ldots, X_n$ is a random sample from this population, and
- assume that one of the following conditions holds:
  - $n < 30$ and the population is normal, or
  - $n \geq 30$.
- If σ is unknown, the decision rule is

$$\text{Reject } \mathrm{H}_0 \text{ if } |T| \geq t_{\alpha/2}(n-1),$$

  where T is the test statistic

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}.$$

- This is a two-tailed test.

- For the testing problem

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0,$$

  reject H_0 if $T > t_\alpha(n-1)$.
- For the testing problem

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu < \mu_0,$$

  reject H_0 if $T < -t_\alpha(n-1)$.
- These are one-tailed tests.
- If σ is known, replace s with σ and $t_\alpha(n-1)$ with z_α.
- Recall that if n is large, $t_\alpha(n-1)$ can also be replaced by z_α as an approximation.

Outline

Test Statistics

- Recall that a *parameter* is a number (or vector etc.) related to the population. For example, the population mean μ and population variance σ^2 are parameters. Parameters are not random.
- A *statistic* is a number (or vector etc.) whose value is based on a *random sample*. For example, the sample mean $\overline{X}$ and sample variance S^2 are statistics. Statistics are random variables.
- Most decision rules for hypothesis tests are based on *test statistics*.
- For example, the decision rules from the previous section were based on the test statistic

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}.$$

Once a sample has been collected and a test statistic has been calculated, the p-value of the test can also be calculated.

Definition

The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming the null hypothesis is true.

Outline

- Let p be a population proportion, and consider the testing problem

$$\mathrm{H}_0 : p = p_0 \text{ vs. } \mathrm{H} : p \neq p_0.$$

- Suppose $\hat{p}$ is the sample proportion from a sample of size n, and
- assume that both $n\hat{p} \geq 5$ and $n(1 - \hat{p}) \geq 5$.
- Decision rule:

$$\text{Reject } \mathrm{H}_0 \text{ if } |Z| \geq z_{\alpha/2},$$

where Z is the test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}.$$

- Use appropriate modifications for one-tailed tests.

- Let p_1 and p_2 be two population proportions, and consider

$$\mathrm{H}_0 : p_1 = p_2 \text{ vs. } \mathrm{H}_1 : p_1 \neq p_2.$$

- Let $\hat{p}_1 = Y_1/n_1$ and $\hat{p}_2 = Y_2/n_2$ be corresponding sample proportions based on **independent** samples of sizes n_1 and n_2, respectively.
- Also, assume that both $n_i\hat{p}_i \geq 5$ and $n_i(1 - \hat{p}_i) \geq 5$, for $i = 1, 2$.
- Decision rule:

$$\text{Reject } \mathrm{H}_0 \text{ if } |Z| \geq z_{\alpha/2}, \text{ where}$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ and}$$

$$\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}.$$

Outline

- Let $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ be two **normal** populations,
- with the **same, unknown variance** $\sigma = \sigma_X = \sigma_Y$, and consider

$$H_0 : \mu_X = \mu_Y \text{ vs. } H_1 : \mu_X \neq \mu_Y.$$

- Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be **independent** samples from these populations.
- Decision rule:

$$\text{Reject } H_0 \text{ if } |T| \geq t_{\alpha/2}(n+m-2), \text{ where}$$

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}, \text{ and}$$

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

- This test is valid for **any sample sizes**, but it's only necessary to use this test for **small sample sizes**.

- Let $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ be two **normal** populations,
- with **known variances** σ_X and σ_Y, and consider

$$H_0 : \mu_X = \mu_Y \text{ vs. } H_1 : \mu_X \neq \mu_Y.$$

- Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be **independent** samples from these populations.
- Decision rule:

$$\text{Reject } H_0 \text{ if } |Z| \geq z_{\alpha/2}, \text{ where}$$

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}.$$

- This test is valid for **any sample sizes.**

- Consider two populations with means μ_X and μ_Y
- and variances σ_X and σ_Y. Consider

$$\mathrm{H}_0 : \mu_X = \mu_Y \text{ vs. } \mathrm{H}_1 : \mu_X \neq \mu_Y.$$

- Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be **independent** samples from these populations, where
- $n \geq 30$ and $m \geq 30$.
- Decision rule:

$$\text{Reject } \mathrm{H}_0 \text{ if } |Z| \geq z_{\alpha/2}, \text{ where}$$

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}.$$

- This test is **only** valid for **large sample sizes.**
- If the population variances are known, use them in Z instead of the sample variances.

- Consider a sample of **paired observations** $(X_1, Y_1), \ldots, (X_n, Y_n)$ and

$$\mathrm{H}_0 : \mu_X = \mu_Y \text{ vs. } \mathrm{H}_1 : \mu_X \neq \mu_Y.$$

- Define the differences $D_i = X_i - Y_i$, for $i = 1, \ldots, n$.
- The above testing problem is now equivalent to testing

$$\mathrm{H}_0 : \mu_D = 0 \text{ vs. } \mathrm{H}_1 : \mu_D \neq 0,$$

  and can be solved using the techniques from Section 7.2.
- Decision rule:

$$\text{Reject } \mathrm{H}_0 \text{ if } |T| \geq t_{\alpha/2}(n-1), \text{ where}$$

$$T = \frac{\overline{D}}{S_D/\sqrt{n}}.$$

- This test is only valid if the populations are **normal or $n \geq 30$.**

Outline

- Let $\mathrm{N}(\mu_X, \sigma_X^2)$ be a **normal population**, and consider

$$\mathrm{H}_0 : \sigma_X^2 = \sigma_0^2 \text{ vs. } \mathrm{H}_1 : \sigma_X^2 \neq \sigma_0^2.$$

- Let $X_1, \ldots, X_n$ be a random sample from this population.
- Decision rule:

$$\text{Reject } \mathrm{H}_0 \text{ if } \chi^2 \leq \chi^2_{1-\alpha/2}(n-1) \text{ or } \chi^2 \geq \chi^2_{\alpha/2}(n-1), \text{ where}$$

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}.$$

- This test is valid for **any sample size.**
- This test is not robust against deviations from normality.

- Let $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$ be two **normal** populations, and consider

$$H_0 : \sigma_X^2 = \sigma_Y^2 \text{ vs. } H_1 : \sigma_X^2 \neq \sigma_Y^2.$$

- Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be **independent** samples from these populations.
- Decision rule:

Reject H_0 if $F \leq F_{1-\alpha/2}(n-1, m-1)$ or $F \geq F_{\alpha/2}(n-1, m-1)$, where

$$F = \frac{S_X^2}{S_Y^2}.$$

- This test is valid for **any sample sizes.**
- This test is not robust against deviations from normality.

Outline

Simple Linear Regression

- A *simple linear regression model* is given by

$$Y_i = a + bx_i + \epsilon_i, \text{ for } i = 1, \ldots, n, \text{ where}$$

- $Y_1, \ldots, Y_n$ and $x_1, \ldots, x_n$ are measurements of some physical quantities.
- We are interested in finding a relationship between these quantities and using x to predict Y.
- The Y_i's are **random variables**, while the x_i's can be regarded as **random or constant.**
- Both the Y_i's and x_i's **are observable**.
- x is often called the **independent, explanatory, or input variable**.
- Y is often called the **dependent, response, or output variable**.

Simple Linear Regression

- $$Y_i = a + bx_i + \epsilon_i, \text{ for } i = 1, \ldots, n, \text{ where}$$
- This model assumes there is a **linear relationship** between x and Y, except for random error.
- The **parameters** a and b are the y-intercept and slope of this line.
- a and b are fixed **numbers.**
- a and b are **not observable.**

Simple Linear Regression

- $$Y_i = a + bx_i + \epsilon_i, \text{ for } i = 1, \ldots, n, \text{ where}$$
- The ϵ_i's are **random errors.**
- We will assume that $\epsilon_1, \ldots, \epsilon_n$ are IID $N(0, \sigma^2)$ **random variables.**
- σ^2 is a **parameter** (it's just a number, not random).
- The ϵ_i's and σ^2 are **not observable.**

Simple Linear Regression: Concise Formulation

- A *simple linear regression model* is given by

$$Y_i = a + bx_i + \epsilon_i, \text{ for } i = 1, \ldots, n, \text{ where}$$

- $Y_1, \ldots, Y_n$ are observable random variables;
- $x_1, \ldots, x_n$ are observable numbers;
- a and b are unobservable numbers;
- $\epsilon_1, \ldots, \epsilon_n$ are unobservable IID $N(0, \sigma^2)$ random variables; and
- σ^2 is an unobservable positive number.

Hook's Law

- Hook's Law from physics states that the length of a spring is a linear function of the mass placed on the spring.
- Consider the following data.

Mass (kg)	Length (cm)
0	439.00
2	439.12
4	439.21
6	439.31
8	439.40
10	439.50

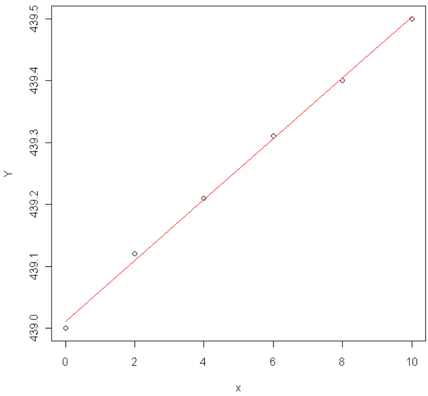
- Does the data fall exactly on a line? Why not?
- Find the slope and y-intercept of the best fitting line.

Estimating the y-intercept and Slope

- For the simple linear regression model, the MLEs for a and b are **least squares estimates**: they minimize the sum of the squares of the distances from the data points to the regression line.
- $$\hat{b} = r\frac{S_Y}{S_x} = \frac{\sum x_i y_i - 1/n(\sum x_i)(\sum y_i)}{\sum x_i^2 - 1/n(\sum x_i)^2}.$$
- The regression line passes through the vector of means $(\overline{x}, \overline{Y})$. Therefore,
$$\hat{a} = \overline{Y} - \hat{b}\overline{x}.$$
- Note that $\hat{a}$ and $\hat{b}$ are observable random variables.

Estimates for Hook's Law Example

- For the Hook's Law example, $\hat{b} = 0.0491$ and $\hat{a} = 439.01$
- Scatter plot and least squares line:



- How would we estimate σ^2?

Residuals and Estimating σ^2

- Rearranging the regression equation, we obtain

$$\epsilon_i = Y_i - (a + bx_i), \text{ for } i = 1, \ldots, n.$$

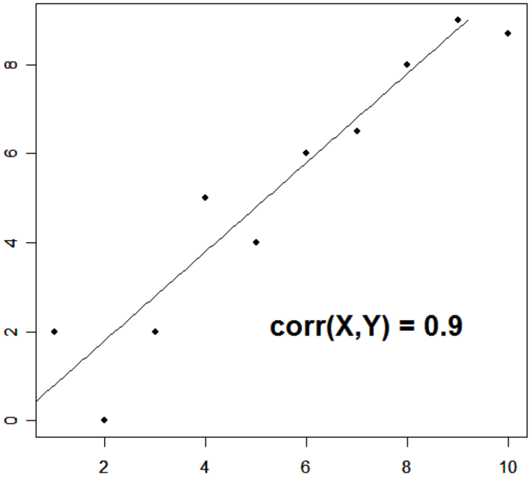
- We can't directly observe the ϵ_i's.
- We can observe the **residuals**, given by the equation below:

$$e_i = Y_i - (\hat{a} + \hat{b}x_i), \text{ for } i = 1, \ldots, n.$$

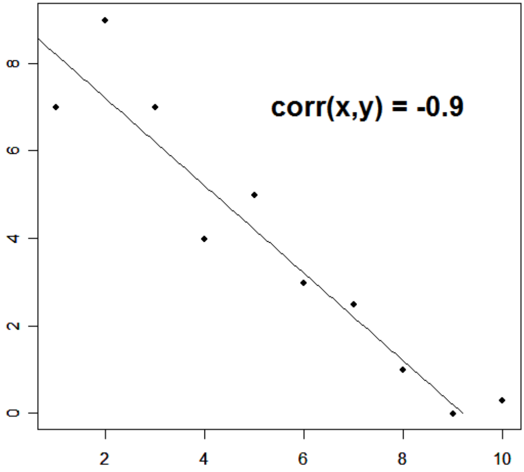
- The residuals are observable random variables.
- If our estimates are accurate, $e_i \approx \epsilon_i$.
- The (unbiased) MLE for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2.$$

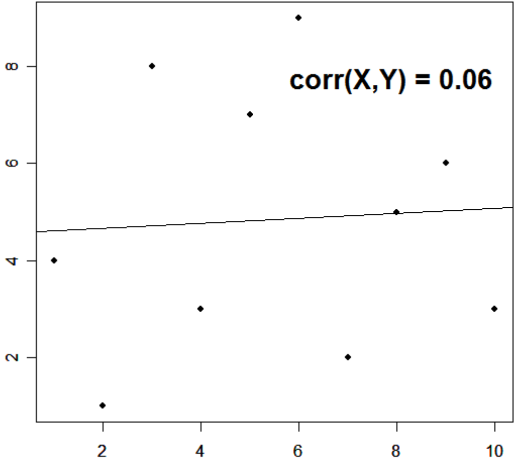
Strong Positive Correlation



Strong Negative Correlation



Virtually No Correlation



Multiple Regression

- $$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \text{ for } i = 1, \ldots, n, \text{ where}$$
- $\epsilon_1, \ldots, \epsilon_n$ are unobservable IID $N(0, \sigma^2)$ random variables.
- $$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$
  $$Y = X\beta + \epsilon$$
- X is called the *design matrix*.

- $$Y = X\beta + \epsilon$$
- $$\hat{\beta} = (X^t X)^{-1} X^t Y$$
- $$\text{cov}(\hat{\beta}) = \sigma^2 (X^t X)^{-1}$$
- The standard error of $\hat{\beta}_j$ is the square root of the jth diagonal entry of $\text{cov}(\hat{\beta})$.
$$\hat{\text{SE}}(\hat{\beta}_j) = \hat{\sigma}\sqrt{[(X^t X)^{-1}]_{jj}}$$
- The following random variable has a t-distribution with $n - p$ degrees of freedom.
$$T = \frac{\hat{\beta}_j - \beta_j}{\hat{\text{SE}}(\hat{\beta}_j)}.$$

- The following random variable has a t-distribution with $n-p$ degrees of freedom.
$$T = \frac{\hat{\beta}_j - \beta_j}{\hat{\text{SE}}(\hat{\beta}_j)}.$$
- $1-\alpha$ confidence interval of β_j:
$$\hat{\beta}_j \pm t_{\alpha/2}(n-p)\hat{\text{SE}}(\hat{\beta}_j)$$
- To test $\text{H}_0 : \beta_j = 0$ vs. $\text{H}_1 : \beta_j \neq 0$, compute
$$T = \frac{\hat{\beta}_j}{\hat{\text{SE}}(\hat{\beta}_j)},$$
and reject if $|T| \geq t_{\alpha/2}(n-p)$.

The F-test

- Consider the multiple regression model

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

- Let p_0 be a positive integer such that $1 \leq p_0 \leq p$.
- Consider testing the null hypothesis that the first p_0 components in the regression equation are zero, i.e.,

$$\mathrm{H}_0 : \beta_1 = \cdots = \beta_{p_0} = 0 \text{ vs. } \mathrm{H}_1 : \beta_j \neq 0, \text{ for at least one } j = 1, \ldots, p_0.$$

- Let e be the vector of residuals for the full model.
- Let $e^{(s)}$ be the vector of residuals for the small model.
- The F-statistic used for this testing problem is

$$F = \frac{(\|e^{(s)}\|^2 - \|e\|^2)/p_0}{\|e\|^2/(n-p)}.$$

The F-test

-
  $$H_0 : \beta_1 = \cdots = \beta_{p_0} = 0 \text{ vs. } H_1 : \beta_j \neq 0, \text{ for at least one } j = 1, \ldots, p_0.$$
- The F-statistic used for this testing problem is
  $$F = \frac{(\|e^{(s)}\|^2 - \|e\|^2)/p_0}{\|e\|^2/(n-p)}.$$
- Under the null hypothesis, F has an $F(p_0, n-p)$ distribution.