Week 6
Nonparametrics

Bodhisattva Sen

December 6, 2017

1 The sample distribution function

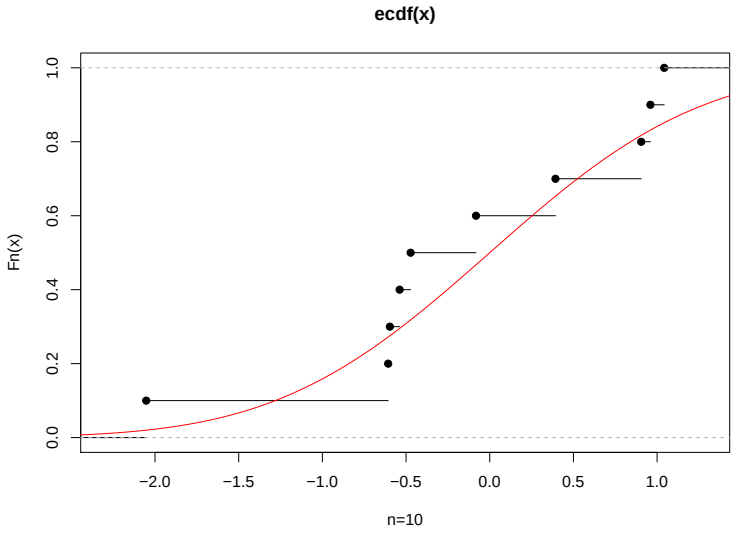
Let $X_1, \ldots, X_n$ be i.i.d F, where F is an unknown distribution function.

Question: We want to estimate F without assuming any specific parametric form for F.

Empirical distribution function (EDF): For each $x \in \mathbb{R}$, we define $F_n(x)$ as the proportion of observed values in the sample that are less than of equal to x, i.e.,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i).$$

The function F_n defined in this way is called the *sample/empirical distribution function*.



Idea: Note that

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{E}[I_{(-\infty,x]}(X)].$$

Thus, given a random sample, we can find an *unbiased* estimator of $F(x)$ by looking at the proportion of times, among the X_i's, we observe a value $\leq x$.

By the WLLN, we know that

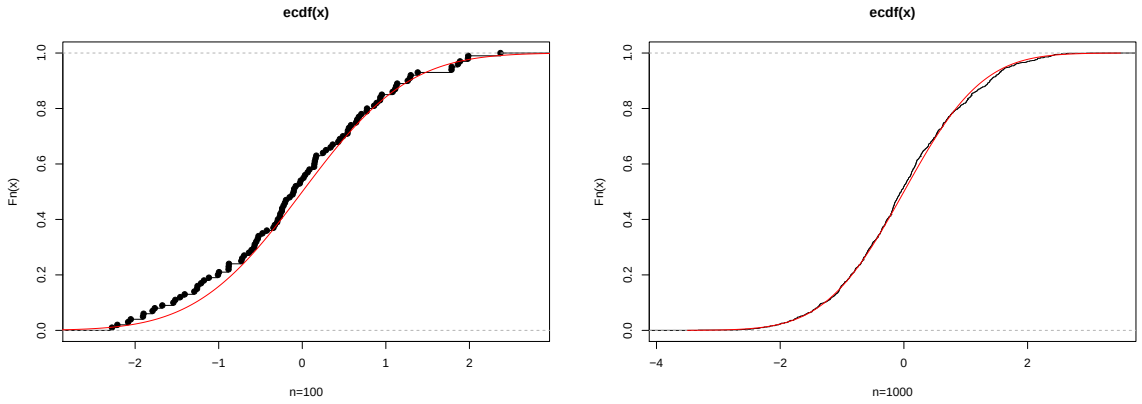
$$F_n(x) \xrightarrow{p} F(x), \qquad \text{for every } x \in \mathbb{R}.$$

Theorem 1. Glivenko-Cantelli Theorem. *Let F_n be the sample c.d.f from an i.i.d sample $X_1, \ldots, X_n$ from the c.d.f F. Then,*

$$D_n := \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{p} 0.$$

By the CLT, we have

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x))), \qquad \text{for every } x \in \mathbb{R}.$$



As $F_n(x) \xrightarrow{p} F(x)$ for all $x \in \mathbb{R}$, we can also say that

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F_n(x)(1 - F_n(x))}} \xrightarrow{d} N(0, 1), \qquad \text{for every } x \in \mathbb{R}.$$

Thus, an asymptotic $(1 - \alpha)$ CI for $F(x)$ is

$$\left[F_n(x) - \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{F_n(x)(1 - F_n(x))}, F_n(x) + \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{F_n(x)(1 - F_n(x))} \right].$$

Likewise, we can also test the hypothesis $H_0 : F(x) = F_0(x)$ versus $H_1 : F(x) \neq F_0(x)$ for some known fixed c.d.f F_0, and $x \in \mathbb{R}$.

1.1 The Kolmogorov-Smirnov goodness-of-fit test

Suppose that we wish to test the simple null hypothesis that the unknown c.d.f F is actually a particular continuous c.d.f F^* against the alternative that the actual c.d.f is not F^*, i.e.,

$$H_0 : F(x) = F^*(x) \quad \text{for } x \in \mathbb{R}, \qquad H_0 : F(x) \neq F^*(x) \quad \text{for some } x \in \mathbb{R}.$$

This is a nonparametric ("infinite" dimensional) problem.

Let

$$D_n^* = \sup_{x \in \mathbb{R}} |F_n(x) - F^*(x)|.$$

D_n^* is the maximum difference between the sample c.d.f F_n and the hypothesized c.d.f F^*.

We should reject H_0 when

$$n^{1/2} D_n^* \geq c_\alpha.$$

This is called the **Kolmogorov-Smirnov** test.

How do we find c_α?

When H_0 is true, the distribution of D_n^* will have a certain distribution that is the same for every possible continuous c.d.f F. (Why?)

Note that, under H_0,

$$\begin{aligned}
D_n^* &= \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n I(X_i \leq x) - F^*(x) \right| \\
&= \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n I(F^*(X_i) \leq F^*(x)) - F^*(x) \right| \\
&= \sup_{F^*(x) \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n I(U_i \leq F^*(x)) - F^*(x) \right| = \sup_{t \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^n I(U_i \leq t) - t \right| \\
&= \sup_{t \in [0,1]} |F_{n,U}(t) - t| ,
\end{aligned}$$

where $U_i := F^*(X_i) \sim \text{Uniform}(0,1)$ (i.i.d) and $F_{n,U}$ is the EDF of the U_i's. Thus, D_n^* is *distribution-free*.

Theorem 2. *(Distribution-free property) Under H_0, the distribution of D_n^* is the same for all continuous distribution functions F.*

We also have the following theorem.

Theorem 3. *Under H_0, as $n \to \infty$,*

$$n^{1/2} D_n^* \xrightarrow{d} H, \tag{1}$$

where H is a valid c.d.f.

In fact, the exact sampling distribution of the KS statistic, under H_0, can be approximated by *simulations*, i.e., we can draw n *data points from a Uniform(0,1) distribution and recompute the test statistic* multiple times.

1.1.1 The Kolmogorov-Smirnov test for two samples

Consider a problem in which a random sample of m observations $X_1, \ldots, X_m$ is taken from the unknown c.d.f F, and an independent random sample of n observations $Y_1, \ldots, Y_n$ is taken from another distribution with unknown c.d.f G.

It is desired to test the hypothesis that both these functions, F and G, are identical, without specifying their common form. Thus the hypotheses we want to test are:

$$H_0 : F(x) = G(x) \quad \text{for } x \in \mathbb{R}, \qquad H_0 : F(x) \neq G(x) \quad \text{for some } x \in \mathbb{R}.$$

We shall denote by F_m the EDF of the observed sample $X_1, \ldots, X_m$, and by G_n the EDF of the sample $Y_1, \ldots, Y_n$.

We consider the following statistic:

$$D_{m,n} = \sup_{x \in \mathbb{R}} |F_m(x) - G_n(x)|.$$

When H_0 holds, the sample EDFs F_m and G_n will tend to be close to each other. In fact, when H_0 is true, it follows from the Glivenko-Cantelli lemma that

$$D_{m,n} \xrightarrow{p} 0 \qquad \text{as } m, n \to \infty.$$

$D_{m,n}$ is also *distribution-free* (why?)

Theorem 4. *Under H_0,*

$$\left(\frac{mn}{m+n}\right)^{1/2} D_{m,n} \xrightarrow{d} H,$$

where H is a the same c.d.f as in (1)*.*

A test procedure that rejects H_0 when

$$\left(\frac{mn}{m+n}\right)^{1/2} D_{m,n} \geq c_\alpha,$$

where c_α (is the $(1-\alpha)$-quantile of H) is an appropriate constant, is called a *Kolmogorov-Smirnov two sample test.*

Exercise: Show that this test statistic is also distribution-free under H_0. Thus, the critical of the test can be obtained via simulations.

2 Bootstrap

Example 1: Suppose that we model our data $\mathbf{X} = (X_1, \ldots, X_n)$ as coming from some distribution with c.d.f F having median θ.

Suppose that we are interested in using the sample median M as an estimator of θ.

We would like to estimate the MSE (mean squared error) of M (as an estimator of θ), i.e., we would like to estimate

$$\mathbb{E}[(M-\theta)^2].$$

We may also be interested in finding a confidence interval for θ.

Example 2: Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from a distribution F. We are interested in the distribution of the sample correlation coefficient:

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2\right]^{1/2}}.$$

We might be interested in the variance of R, or the bias of R, or the distribution of R as an estimator of the correlation ρ between X and Y.

Question: How do we get a handle on these problems?

How would we do it if an *oracle* told us F?

Bootstrap: The bootstrap is a method of replacing (plug-in) an unknown distribution function F with a known distribution in probability/expectation calculations.

If we have a sample of data from the distribution F, we first approximate F by $\hat{F}$ and then perform the desired calculation.

If $\hat{F}$ is a good approximation of F, then bootstrap can be successful.

2.1 Bootstrap in general

Let $\eta(\mathbf{X}, F)$ be a quantity of interest that possibly depends on both the distribution F and a sample $\mathbf{X}$ drawn from F.

In general, we might wish to estimate the mean or a quantile or some other probabilistic feature or the entire *distribution* of $\eta(\mathbf{X}, F)$.

The bootstrap estimates $\eta(\mathbf{X}, F)$ by $\eta(\mathbf{X}^*, \hat{F})$, where $\mathbf{X}^*$ is a random sample drawn from the distribution $\hat{F}$, where $\hat{F}$ is some distribution that we think is close to F.

How do we find the distribution of $\eta(\mathbf{X}^*, \hat{F})$?

In most cases, the distribution of $\eta(\mathbf{X}^*, \hat{F})$ is difficult to compute, but we can approximate it easily by simulation.

The bootstrap can be broken down in the following simple steps:

- Find a "good" estimator $\hat{F}$ of F.
- Draw a large number (say, v) of random samples $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(v)}$ from the distribution $\hat{F}$ and then compute $T^{(i)} = \eta(\mathbf{X}^{*(i)}, \hat{F})$, for $i = 1, \ldots, v$.
- Finally, compute the desired feature of $\eta(\mathbf{X}^*, \hat{F})$ using the sample c.d.f of the values $T^{(1)}, \ldots, T^{(v)}$.

2.2 Parametric bootstrap

Example 1: (Estimating the standard deviation of a statistic)

Suppose that $X_1, \ldots, X_n$ is random sample from $N(\mu, \sigma^2)$.

Suppose that we are interested in the parameter

$$\theta = \mathbb{P}(X \leq c) = \Phi\left(\frac{c - \mu}{\sigma}\right),$$

where c is a given known constant.

What is the MLE of θ?

The MLE of θ is

$$\hat{\theta} = \Phi\left(\frac{c - \bar{X}}{\hat{\sigma}}\right).$$

Question: How do we calculate the standard deviation of $\hat{\theta}$? There is no easy closed form expression for this.

Solution: We can bootstrap!

Draw many (say v) bootstrap samples of size n from $N(\bar{X}, \hat{\sigma}^2)$. For the i-th sample we compute a sample average $\bar{X}^{*(i)}$, a sample standard deviation $\hat{\sigma}^{*(i)}$.

Finally, we compute

$$\hat{\theta}^{*(i)} = \Phi\left(\frac{c - \bar{X}^{*(i)}}{\hat{\sigma}^{*(i)}}\right).$$

We can estimate the mean of $\hat{\theta}$ by

$$\bar{\theta}^* = \frac{1}{v}\sum_{i=1}^{v} \hat{\theta}^{*(i)}.$$

The standard deviation of $\hat{\theta}$ can then be estimated by the sample standard deviation of the $\hat{\theta}^{*(i)}$ values, i.e.,

$$\left[\frac{1}{v}\sum_{i=1}^{v}(\hat{\theta}^{*(i)} - \bar{\theta}^*)^2\right]^{1/2}.$$

Example 2: (Comparing means when variances are unequal) Suppose that we have two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from two possibly different normal populations. Suppose that

$$X_1, \ldots, X_m \text{ are i.i.d } N(\mu_1, \sigma_1^2) \qquad \text{and} \qquad Y_1, \ldots, Y_n \text{ are i.i.d } N(\mu_2, \sigma_2^2).$$

Suppose that we want to test

$$H_0 : \mu_1 = \mu_2 \qquad \text{versus} \qquad H_1 : \mu_1 \neq \mu_2.$$

We can use the test statistic

$$U = \frac{(m+n-2)^{1/2}(\bar{X}_m - \bar{Y}_n)}{\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}(S_X^2 + S_Y^2)^{1/2}}.$$

Note that as $\sigma_1^2 \neq \sigma_2^2$, U does not necessarily follow a t-distribution.

How do we find the cut-off value of the test?

The parametric bootstrap can proceed as follows:

First choose a large number v, and for $i = 1, \ldots, v$, simulate $(\bar{X}_m^{*(i)}, \bar{Y}_n^{*(i)}, S_X^{2*(i)}, S_Y^{2*(i)})$, where all four random variables are independent with the following distributions:

- $\bar{X}_m^{*(i)} \sim N(0, \hat{\sigma}_1^2/m)$.
- $\bar{Y}_n^{*(i)} \sim N(0, \hat{\sigma}_2^2/n)$.
- $S_X^{2*(i)} \sim \hat{\sigma}_1^2\, \chi_{m-1}^2$.

- $S_Y^{2*(i)} \sim \hat{\sigma}_2^2 \, \chi_{n-1}^2$.

Then we compute

$$U^{*(i)} = \frac{(m+n-2)^{1/2}(\bar{X}_m^{*(i)} - \bar{Y}_n^{*(i)})}{\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}(S_X^{2*(i)} + S_Y^{2*(i)})^{1/2}}$$

for each i.

We approximate the null distribution of U by the distribution of the $U^{*(i)}$'s.

Let c^* be the $\left(1-\frac{\alpha}{2}\right)$-quantile of the distribution of $U^{*(i)}$'s. Thus we reject H_0 if

$$|U| > c^*.$$

2.3 The nonparametric bootstrap

Back to Example 1: Let $X_1, \ldots, X_n$ be a random sample from a distribution F.

Suppose that we want a CI for the median θ of F.

We can base a CI on the sample median M.

We want the distribution of $M - \theta$!

Let $\eta(\mathbf{X}, F) = M - \theta$.

We approximate the $\alpha/2$ and the $1-\alpha/2$ quantiles of the distribution of $\eta(\mathbf{X}, F)$ by that of $\eta(\mathbf{X}^*, \hat{F})$.

We may choose $\hat{F} = F_n$, the empirical distribution function. Thus, our method can be broken in the following steps:

- Choose a large number v and simulate many samples $\mathbf{X}^{*(i)}$, for $i = 1, \ldots, n$, from F_n. This reduces to drawing **with replacement sampling** from $\mathbf{X}$.
- For each sample we compute the sample median $M^{*(i)}$ and then find the sample quantiles of $\{M^{*(i)} - M\}_{i=1}^v$.

Back to Example 2: Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from a distribution F. We are interested in the distribution of the sample correlation coefficient:

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2\right]^{1/2}}.$$

We might be interested in the bias of R, i.e., $\eta(\mathbf{X}, \mathbf{Y}, F) = R - \rho$.

Let F_n be the discrete distribution that assigns probability $1/n$ to each of the n data points.

Thus, our method can be broken in the following steps:

- Choose a large number v and simulate many samples from F_n. This reduces to drawing **with replacement sampling** from the original paired data.
- For each sample we compute the sample correlation coefficient $R^{*(i)}$ and then find the sample quantiles of $\{T^{*(i)} = R^{*(i)} - R\}_{i=1}^{v}$.
- We estimate the mean of $R - \rho$ by the average $\frac{1}{n}\sum_{i=1}^{v} T^{*(i)}$.

3 Review

3.1 Statistics

- Estimation: Maximum likelihood estimation (MLE); large sample properties of the MLE; Information matrix; method of moments.
- Consistency of estimators; Mean squared error and its decomposition; unbiased estimation; minimum variance unbiased estimator; sufficiency.
- Bayes estimators: prior distribution; posterior distribution.
- Sampling distribution of an estimator; sampling from a normal distribution; t-distribution.

Exercise: Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution with mean 0 and unknown variance σ^2. Determine the asymptotic distribution of the statistic $T = \left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)^{-1}$.

Solution: We know that X_i^2's are i.i.d with mean $\mathbb{E}(X_1^2) = \sigma^2$ and Var $(X_1^2) = \mathbb{E}(X_1^4) - [\mathbb{E}(X_1^2)]^2 = 2\sigma^4$. Note that X_i^2's have a χ_1^2 distribution. Thus, by the CLT, we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \sigma^2\right) \xrightarrow{d} N(0, 2\sigma^4).$$

Let $g(x) = x^{-1}$. Thus, $g'(x) = -x^{-2}$. Therefore,

$$\sqrt{n}(T - \sigma^{-2}) \xrightarrow{d} N(0, 2\sigma^4 \cdot \sigma^{-8}).$$

Exercise: Consider i.i.d observations $X_1, \ldots, X_n$ where each X_i follows a normal distribution with mean and variance both equal to $1/\theta$, where $\theta > 0$. Thus,

$$f_\theta(x) = \frac{\sqrt{\theta}}{\sqrt{2\pi}} \exp\left[-\frac{(x - \theta^{-1})^2}{2\theta^{-1}}\right].$$

Show that the MLE is one of the solutions to the equation:

$$\theta^2 W - \theta - 1 = 0,$$

where $W = n^{-1}\sum_{i=1}^n X_i^2$. Determine which root it is and compute its approximate variance in large samples.

Solution: We have the log-likelihood (up to a constant) as

$$\ell(\theta) = \frac{n}{2}\log\theta - \frac{\theta}{2}\sum_{i=1}^{n} X_i^2 + n\bar{X} - \frac{n}{2\theta}.$$

Therefore, the score equation is

$$\begin{aligned} & \frac{\partial \ell}{\partial \theta} = \frac{n}{2\theta} - \frac{1}{2}\sum_{i=1}^{n} X_i^2 + \frac{n}{2\theta^2} = 0 \\ \text{i.e.,} \quad & \frac{1}{2\theta} - \frac{1}{2}W + \frac{1}{2\theta^2} = 0 \\ \text{i.e.,} \quad & W\theta^2 - \theta - 1 = 0 \end{aligned}$$

The two roots are given by

$$\frac{1 \pm \sqrt{1+4W}}{2W}$$

and the admissible root is

$$\hat{\theta}_{MLE} = \frac{1+\sqrt{1+4W}}{2W}.$$

We know that

$$\hat{\theta}_{MLE} \sim N\left(\theta, \frac{1}{nI(\theta)}\right) \qquad \text{(approximately)}.$$

Thus the approximate variance of $\hat{\theta}_{MLE}$ is $\frac{1}{nI(\theta)}$, where

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X_1)\right] = \frac{1}{2\theta^2} + \frac{1}{\theta^3}.$$

- Confidence intervals; Cramer-Rao information inequality.

Exercise: A biologist is interested in measuring the ratio of mean weight of animals of two species. However, the species are extremely rare and after much effort she succeeds in measuring the weight of one animal from the first species and one from the second. Let X_1 and X_2 denote these weights. It is assumed that $X_i \sim N(\theta_i, 1)$, for $i = 1, 2$. Interest lies in estimating θ_1/θ_2.

Compute the distribution of

$$h(X_1, X_2, \theta_1, \theta_2) = \frac{\theta_2 X_1 - \theta_1 X_2}{\sqrt{\theta_1^2 + \theta_2^2}}.$$

Is

$$\frac{X_1 - (\theta_1/\theta_2)X_2}{\sqrt{(\theta_1/\theta_2)^2 + 1}}$$

a pivot? Discuss how you can construct a confidence set for the ratio of mean weights.

Solution: Note that $\theta_2 X_1 - \theta_1 X_2 \sim N(0, \theta_1^2 + \theta_2^2)$ as

$$\mathbb{E}(\theta_2 X_1 - \theta_1 X_2) = \theta_2\theta_1 - \theta_1\theta_2 = 0$$

and

$$\mathrm{Var}\,(\theta_2 X_1 - \theta_1 X_2) = \mathrm{Var}\,(\theta_2 X_1) + \mathrm{Var}\,(\theta_1 X_2) = \theta_2^2 + \theta_1^2.$$

Thus,

$$h(X_1, X_2, \theta_1, \theta_2) = \frac{\theta_2 X_1 - \theta_1 X_2}{\sqrt{\theta_1^2 + \theta_2^2}} \sim N(0,1).$$

Now,

$$\frac{X_1 - (\theta_1/\theta_2)X_2}{\sqrt{(\theta_1/\theta_2)^2 + 1}} = \frac{\theta_2 X_1 - \theta_1 X_2}{\sqrt{\theta_1^2 + \theta_2^2}} \sim N(0,1)$$

and is thus indeed a pivot.

To get a confidence set for $\eta := \theta_1/\theta_2$, we know that

$$\mathbb{P}\left[-z_{\alpha/2} \le \frac{X_1 - \eta X_2}{\sqrt{\eta^2 + 1}} \le z_{\alpha/2}\right] = 1 - \alpha$$

$$\text{i.e.,} \quad \mathbb{P}\left[\frac{|X_1 - \eta X_2|}{\sqrt{\eta^2 + 1}} \le z_{\alpha/2}\right] = 1 - \alpha$$

$$\text{i.e.,} \quad \mathbb{P}\left[(X_1 - \eta X_2)^2 - (\eta^2 + 1)^2 z_{\alpha/2} \le 0\right] = 1 - \alpha.$$

Thus,

$$\{\eta : (X_1 - \eta X_2)^2 - (\eta^2 + 1)^2 z_{\alpha/2} \le 0\}$$

gives a level $(1 - \alpha)$ confidence set for η. This can be expressed explicitly in terms of the roots of the quadratic equation involved.

- Hypothesis testing: Null and the alternative hypothesis; rejection region; Type I and II errorx; power function; size (level) of a test; equivalence of tests and confidence sets; p-value; Neyman-Pearson lemma; uniformly most powerful test.
- t-test; F-test; likelihood ratio test
- Linear models: method of least squares; regression; Simple linear regression; inference on β_0 and β_1; mean response; prediction interval;
- General linear model; MLE; projection; one-way ANOVA

Exercise: Processors usually preserve cucumbers by fermenting them in a low-salt brine (6% to 9% sodium chloride) and then storing them in a high-salt brine until they are used by processors to produce various types of pickles. The high-salt brine is needed to retard softening of the pickles and to prevent freezing when they are stored outside in northern climates. Data showing the reduction in firmness of pickles stored over time in a low-salt brine (2% to 3%) are given in the following table.

Weeks (X) in Storage at 72^o F	0	4	14	32	52
Firmness (Y) in pounds	19.8	16.5	12.8	8.1	7.5

(a) Fit a least-squares line to the data.

(b) Compute R^2 to evaluate the goodness of the fit to the data points?

(c) Use the least-squares line to estimate the mean firmness of pickles stored for 20 weeks.

(d) Determine the 95% CI for β_1.

(e) Test the null hypothesis that Y does not depend on X linearly.

Solution: (a) Fit a least-squares line to the data.

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{-425.48}{1859.2} = -0.229$$
$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x} = 12.94 - (-0.229)(20.4) = 17.612$$
$$\widehat{y} = 17.612 - 0.229x.$$

(b) Compute R^2 to evaluate the goodness of the fit to the data points?

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{15.4}{112.772} = 0.863.$$

(c) Use the least-squares line to estimate the mean firmness of pickles stored for 20 weeks.

$$\widehat{y}(20) = 17.612 - (0.229)(20) = 13.0$$

(d) Determine the 95% CI for β_1.

The 95% CI for β_1 is given by

$$\widehat{\beta_1} \pm t_{0.025,3} SE(\widehat{\beta_1}) \quad \text{where} \quad SE(\widehat{\beta_1}) = \frac{s}{\sqrt{S_{xx}}}.$$

We have $s = \sqrt{SSE/3} = \sqrt{(15.4)/3} = 2.266$, thus $SE(\widehat{\beta}_1) = \frac{2.66}{\sqrt{1859.2}} = 0.052$. Thus the 95% CI for β_1 is

$$-0.229 \pm (3.18)(0.052) = [-0.396, -0.062]$$

(e) Test the null hypothesis that Y does not depend on X linearly.

We test the hypothesis

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0$$

with level at $\alpha = 0.05$. This can be tested with t-statistic

$$T = \frac{\widehat{\beta}_1}{SE(\widehat{\beta}_1)} \quad \text{and} \quad RR : |t| > t_{0.025,3} = 3.18.$$

The observed $t = \dfrac{-0.229}{0.052} = -4.404$, which is in the rejection region. Thus we reject the hypothesis that $\beta_1 = 0$. This means based on the data we reject H_0.

Exercise: A manager wishes to determine whether the mean times required to complete a certain task differ for the three levels of employee training. He randomly selected 10 employees with each of the three levels of training (Beginner, Intermediate and Advanced). Do the data provide sufficient evidence to indicate that the mean times required to complete a certain task differ for at least two of the three levels of training? The data is summarized in the following table. Use the level $\alpha = 0.05$.

	$\overline{x_i}$	s_i^2
Advanced	24.2	21.54
Intermediate	27.1	18.64
Beginner	30.2	17.76

Solution: Let α_i denote the mean effect of ith training level; advanced=1, intermediate=2 and beginner=3. We test the hypothesis

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 \quad \text{vs.} \quad H_a : \alpha_i \neq \alpha_j \quad \text{for some } i \text{ and } j$$

We have

$$\begin{aligned}
\overline{x_{1\cdot}} &= 24.2 \qquad \overline{x}_{2\cdot} = 27.1 \qquad \overline{x}_{3\cdot} = 30.2 \\
\overline{x}_{\cdot\cdot} &= \frac{1}{3}(24.2 + 27.1 + 30.2) = 27.17 \\
SSB &= 10\left((24.2 - 27.17)^2 + (27.1 - 27.17)^2 + (30.2 - 27.17)^2\right) = 180.1 \\
SSW &= 9(21.54 + 18.64 + 17.76) = 521.46
\end{aligned}$$

Thus we have the following ANOVA-table:

Source of variations	df	SS	MS	F
Treatments	2	180.1	90.03	4.67
Errors	27	521.46	19.31	
Total	29	683.52		

Since the observed $f = 4.67$ is in $RR : f > f_{0.05,2,27} = 3.35$, we reject the H_0. Thus the levels of training appear to have different effects on the mean times required to complete the task.

- The empirical distribution function; goodness-of-fit-tests; Kolmogorov-Smirnov tests.

Read the following sections from the text book for the final:

- Chapter 6 excluding 6.4
- Chapter 7 excluding 7.8, 7.9
- Chapter 8 excluding 8.6
- Chapter 9 excluding 9.3, 9.8, 9.9
- Chapter 10 only 10.6
- Chapter 11 excluding 11.4, 11.7, 11.8

Thank you!

Please complete course evaluations!

Questions?