Examples and Homework

The following data sets will be used for examples and homework.

Data Set 1: 13.1, 12.2, 13.0, 12.9, 13.4, 14.1, 13.8, 12.8, 12.7, 13.3, 14.1, 13.6

The example problems use Data Set 1. The homework problems use Data Set 2.

Sample mean, Sample variance, Central limit theorem

Let \( x_1, \ldots, x_n \) be sample values from iid random variables\(^1\) having mean \( \mu \) and variance \( \sigma^2 \).

The sample mean is

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

The sample mean is an estimate for \( \mu \). Moreover, \( E\bar{x} = \mu \), thus \( \bar{x} \) is unbiased (an estimate is unbiased if its expectation equals that which is estimated).

The sample variance is

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

The sample variance is an estimate for \( \sigma^2 \). Moreover, \( E s^2 = \sigma^2 \) (thus \( s^2 \) is unbiased).

For large \( n \), the following approximation holds (“central limit theorem”)

\[
\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)
\]

(1)

If the samples \( x_i \) come from a normal distribution, then

\[
\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}
\]

(2)

where \( \xi \sim t_k \) has a “Student’s t” distribution with \( k \) degrees of freedom and density

\[
\rho_\xi(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}{\Gamma\left(\frac{k}{2}\right)}}\left(1 + \frac{x^2}{k}\right)^{-(k+1)/2}
\]

The standard error of \( \bar{x} \) is its standard deviation

\[
s.e.(\bar{x}) = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}
\]

(more generally, the standard error of a sample of size \( n \) is often taken to be the sample’s standard deviation divided by \( \sqrt{n} \)).

If the samples \( x_i \) come from a Bernoulli \( p \) distribution, then \( \bar{x} \) estimates the probability \( p \). The estimate is unbiased \( (E\bar{x} = p) \) and is sometimes denoted by \( \bar{p} \). Its standard error is typically estimated as

\[
s.e.(\bar{p}) \approx \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}
\]

\(^1\)Independent random variables are said to be iid if they are identically distributed (have the same distribution).
Examples (Data Set 1)

Glass fiber reinforced polymer specimens were placed in a tension testing machine. Increasing amounts of stress were applied until failure, and the maximum loads are shown in Data Set 1.

1. Estimate the expected failure load.

\[
\bar{x} = \frac{13.1 + 12.2 + 13.0 + 12.9 + 13.4 + 14.1 + 13.8 + 12.8 + 12.7 + 13.3 + 14.1 + 13.6}{12} = 13.25
\]

2. Estimate the variance in failure load.

\[
\approx 0.33727
\]

3. Estimate the probability that the failure load is less than 13.0.

Replace each \(x_i\) in Data Set 1 with \([x_i < 13.0]\) to obtain Bernoulli \(p\) sample values (1 ↔ success, 0 ↔ failure) and use \(\bar{x}\) to estimate the probability.

\[
\bar{x} = \frac{[13.1 < 13.0] + [12.2 < 13.0] + [13.0 < 13.0] + [12.9 < 13.0] + [13.4 < 13.0] + [14.1 < 13.0] + [13.8 < 13.0] + [12.8 < 13.0] + [12.7 < 13.0] + [13.3 < 13.0] + [14.1 < 13.0] + [13.6 < 13.0]}{12} = 1/3
\]

Parameter estimation

Let \(x_1, \ldots, x_n\) be sample values from \(iid\) random variables having mean \(\mu\) and probability distribution depending upon a parameter \(\theta\). The “method of moments” point estimate of \(\theta\) is the solution to \(\bar{x} = \mu\), where \(\mu\) is the distribution’s mean.\(^2\)

If the probability distribution depends on parameters \(\theta_1, \theta_2\), then the “method of moments” point estimate of \(\theta_1, \theta_2\) is the solution to \(\bar{x} = \mu, s^2 = \sigma^2\), where \(\mu\) and \(\sigma^2\) are the distribution’s mean and variance (respectively).\(^3\)

\(^2\)Assuming \(\bar{x} = \mu\) determines \(\theta\).

\(^3\)Assuming \(\bar{x} = \mu, s^2 = \sigma^2\) determine \(\theta_1\) and \(\theta_2\).
If the probability distribution depends on parameters $\theta_1, \ldots, \theta_k$, then the **Maximum Likelihood Estimate** of $\theta_1, \ldots, \theta_k$ is given by parameter values maximizing the following **likelihood function**

$$
\prod_{i=1}^{n} f(x_i)
$$

where $f$ is the probability density (in the “continuous case”) or else the probability mass function (in the discrete case).

**Examples (Data Set 1)**

4. Failure times (in minutes) for over-stressed components are shown in Data Set 1. Unknown to the experimenter, failure times are described by the random variable $\xi + c$ where $c$ is a constant and $\xi \sim \text{Exponential} \lambda$. To approximate the probability an over-stressed component will function at least 15 minutes, the experimenter tests 100 components and counts how many last that long. Estimate the probability that the experimenter’s approximation will be within 0.01 of the correct answer. Estimate that probability if 500 components are tested.

**Estimates for the expectation and variance of failure times**

$$
\mathcal{E}(\xi + c) = c + \mathcal{E}\xi = c + 1/\lambda, \quad \text{var}(\xi + c) = \text{var}(\xi) = 1/\lambda^2
$$

are given by $\bar{x} = 13.25$, $s^2 \approx 0.33727$. Solving gives $c \approx 12.67$, $\lambda \approx 1.722$. Using these values, an over-stressed component will function at least 15 minutes with probability

$$
p = P(\xi + c \geq 15) = \int [x + c \geq 15] \lambda e^{-\lambda x} \, dx \approx 0.0181
$$

The experimenter’s approximation is a sample value of the random variable

$$
\frac{n}{100} = \frac{1}{100} \sum_{i=1}^{100} \left[ \xi_i + c \geq 15 \right] \quad (\eta \sim \text{Binomial}_{100, p})
$$

Using the value $p = 0.0181$, the experimenter’s approximation will be within 0.01 with probability

$$
P\left( \left| \frac{n}{100} - p \right| \leq 0.01 \right) = \sum_{x=0}^{100} \left[ \left| \frac{x}{100} - p \right| \leq 0.01 \right] \left(100 \atop x\right) p^x (1-p)^{100-x} \approx 0.57
$$

If 500 components are tested,

$$
\frac{n}{500} = \frac{1}{500} \sum_{i=1}^{500} \left[ \xi_i + c \geq 15 \right] \quad (\eta \sim \text{Binomial}_{500, p})
$$

and the approximation becomes

$$
P\left( \left| \frac{n}{500} - p \right| \leq 0.01 \right) = \sum_{x=0}^{500} \left[ \left| \frac{x}{500} - p \right| \leq 0.01 \right] \left(500 \atop x\right) p^x (1-p)^{500-x} \approx 0.91
$$

---

4 Assuming that maximizing (3) determines $\theta_1, \ldots, \theta_k$. 

3
5. Suppose \( \xi \sim \text{Multinomial}_{n, p_1, \ldots, p_k} \). Find maximum likelihood estimates (MLE) for the \( p_i \).

Interpreting Data Set 1 as a list of 4 triples, compute the MLE for \( p_1, p_2, p_3 \) \((n = 4, k = 3)\), assuming appropriate units (hundreds, for example) so that the data is integer-valued.

The likelihood function is
\[
\ell = \prod_{x \in D} (x_1, \ldots, x_k)^n p_1^{x_1} \cdots p_k^{x_k}
\]

Maximizing the above coincides with maximizing its logarithm subject to the constraint \( 1 = p_1 + \cdots + p_k \) (the probabilities should sum to 1). Using the method of Lagrange multipliers,
\[
\lambda = \lambda \frac{\partial}{\partial p_j} \sum_i p_i = \frac{\partial}{\partial p_j} \log \ell = \sum_{x \in D} \frac{\partial}{\partial p_j} x_j \log p_j = \sum_{x \in D} x_j / p_j
\]
\[
1 = p_1 + \cdots + p_k
\]
Thus
\[
p_j = \lambda^{-1} \sum_{x \in D} x_j, \quad \lambda = \sum_j \sum_{x \in D} x_j
\]

Using Data Set 1 \((n = 4, k = 3)\),
\[
\lambda = (13.1 + 12.9 + 13.8 + 13.3) + (12.2 + 13.4 + 12.8 + 14.1) + (13.0 + 14.1 + 12.7 + 13.6) = 159
\]
\[
p_1 = (13.1 + 12.9 + 13.8 + 13.3)/159 \approx 0.334
\]
\[
p_2 = (12.2 + 13.4 + 12.8 + 14.1)/159 \approx 0.330
\]
\[
p_3 = (13.0 + 14.1 + 12.7 + 13.6)/159 \approx 0.336
\]

Confidence intervals

A confidence interval \( I \) for \( \theta \) is an interval containing plausible values of \( \theta \); its associated confidence level is the probability that \( \theta \in I \).

Let \( x_1, \ldots, x_n \) be sample values from iid random variables having mean \( \mu \). The most common confidence interval \( I \) for \( \mu \) is to assume (2) and take
\[
I = (\overline{x} - c \frac{s}{\sqrt{n}}, \overline{x} + c \frac{s}{\sqrt{n}})
\]
where \( c > 0 \) is a parameter; \( I \) is often referred to as a two-sided \( t \)-interval\(^5\). The confidence level is \( P(|\xi| < c) \) where
\[
\xi = \frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}
\]
Therefore,
\[
P(|\xi| < c) = 2 \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma((n-1)/2)} \int_0^c \left(1 + \frac{x^2}{n-1}\right)^{-n/2} dx
\]

\(^5\)A \( t \)-interval is sometimes referred to as a variance unknown interval.
A one-sided \( t \)-interval refers to the case where either the left side of \( I \) is \(-\infty \) or else the right side of \( I \) is \( +\infty \). In this case — since the density is an even function — the confidence level is

\[
P(\xi < c) = \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma((n-1)/2)} \int_{-\infty}^{c} \left(1 + \frac{x^2}{n-1}\right)^{-n/2} dx
\]

where \( c \) can be any real value (Extra Credit: prove it).

If \( x_1, \ldots, x_n \) are sample values from iid random variables having mean \( \mu \) and variance \( \sigma^2 \), then the most common confidence interval \( I \) for \( \mu \) is to assume (1) and take

\[
I = (\bar{x} - c \frac{\sigma}{\sqrt{n}}, \bar{x} + c \frac{\sigma}{\sqrt{n}})
\]

where \( c > 0 \) is a parameter; \( I \) is often referred to as a two-sided \( z \)-interval\(^6\). The confidence level is \( P(|\eta| < c) \) where

\[
\eta = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)
\]

Therefore,

\[
P(|\eta| < c) = 2 \frac{1}{\sqrt{2\pi}} \int_{0}^{c} e^{-\frac{x^2}{2}} dx
\]

A one-sided \( z \)-interval refers to the case where either the left side of \( I \) is \(-\infty \) or else the right side of \( I \) is \( +\infty \). In this case — since the density is an even function — the confidence level is

\[
P(\eta < c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} e^{-\frac{x^2}{2}} dx
\]

where \( c \) can be any real value (Extra Credit: prove it).

Examples (Data Set 1)

6. Data Set 1 contains inductance data for a random sample. What is a 99% two-sided \( t \)-interval for the mean component inductance?

Since \( n = 12 \), solving

\[
0.99 = P(|\xi| < c) = 2 \frac{\Gamma(12/2)}{\sqrt{\pi(12-1)}\Gamma((12-1)/2)} \int_{0}^{c} \left(1 + \frac{x^2}{12-1}\right)^{-12/2} dx
\]

gives \( c \approx 3.1058 \). Using \( \bar{x} = 13.25 \) and \( s^2 = 0.33727 \) gives the confidence interval

\[
(\bar{x} - c \frac{s}{\sqrt{n}}, \bar{x} + c \frac{s}{\sqrt{n}}) \approx (12.73, 13.77)
\]

\(^6\)A \( z \)-interval is sometimes referred to as a variance known interval.
7. Data Set 1 contains capacitance data for a random sample. Using the known value 0.58 for the standard deviation, find one-sided 95% confidence intervals for mean component capacitance.

Solving

\[ 0.95 = P(\eta < c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} e^{-\frac{x^2}{2}} dx \]

gives \( c \approx 1.6449 \). Using \( n = 12 \), \( \bar{x} = 13.25 \), \( \sigma = 0.58 \) gives confidence intervals

\((-\infty, \bar{x} + c \frac{\sigma}{\sqrt{n}}) = (-\infty, 13.53)\)

\((\bar{x} - c \frac{\sigma}{\sqrt{n}}, \infty) = (12.97, \infty)\)

8. Data Set 1 contains yields of a chemical process. What is the confidence level of the confidence interval \((-\infty, 13.5)\) for average process yield?

Using \( n = 12 \), \( \bar{x} = 13.25 \), and \( s^2 = 0.33727 \), solving

\[ 13.5 = \bar{x} + c \frac{s}{\sqrt{n}} \]

gives \( c \approx 1.49122 \), and confidence level

\[ P(|\xi| < c) = \frac{\Gamma\left(\frac{12}{2}\right)}{\sqrt{\pi}(12 - 1\Gamma\left(\frac{12-1}{2}\right)} \int_{-\infty}^{c} \left(1 + \frac{x^2}{12 - 1}\right)^{-12/2} dx \approx 0.92 \]

Hypothesis Testing

A null hypothesis \( H_0 \) for mean \( \mu \) is a default assumption concerning possible values for the mean. The alternative hypothesis \( H_A \) is the negation of the null hypothesis. Common examples include

\[ H_A : \mu \neq \mu_0 \quad \text{versus} \quad H_0 : \mu = \mu_0 \]

which is called two-sided (the alternative hypothesis allows \( \mu \) to be on either side of \( \mu_0 \)), and either

\[ H_A : \mu > \mu_0 \quad \text{versus} \quad H_0 : \mu \leq \mu_0 \quad \text{(for this case, let } H^+ \text{ denote } H_0) \]

or

\[ H_A : \mu < \mu_0 \quad \text{versus} \quad H_0 : \mu \geq \mu_0 \quad \text{(for this case, let } H^- \text{ denote } H_0) \]

which are called one-sided (the alternative hypothesis only allows \( \mu \) to be on one side of \( \mu_0 \)).

**Two-sided t-test**: Suppose \( \mu_0 \) is on the boundary of a two-sided \( t \)-interval having confidence \( q \); the \( p \)-value for the null hypothesis \( \mu = \mu_0 \) is defined as \( p = 1 - q \) (\( H_0 \) places \( \mu \) outside the \( t \)-interval). It follows (see confidence intervals) that

\[ p = P(|\xi| \geq c) = 2 \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)\Gamma\left(\frac{n-1}{2}\right)}} \int_{c}^{\infty} \left(1 + \frac{x^2}{n-1}\right)^{-n/2} dx \]

where \( c = |\bar{x} - \mu_0|/(s/\sqrt{n}) \) (Extra Credit: prove it); \( c \) is referred to as the \( t \)-statistic.
One-sided \( t \)-test: Suppose \( \mu_0 \) is on the boundary of a one-sided \( t \)-interval having confidence \( p \); the \( p \)-value for the null hypothesis (either \( H^+ \) or \( H^- \), whichever places \( \mu \) in the \( t \)-interval) is defined as \( p \). It follows (see confidence intervals) that
\[
p = P(\xi \geq c) = \frac{\Gamma(\frac{n}{2})}{\sqrt{\pi(n-1)}\Gamma(\frac{n-1}{2})} \int_{c}^{\infty} \left(1 + \frac{x^2}{n-1}\right)^{-n/2} dx
\]
where \( c = \pm(\pi - \mu_0)/(s/\sqrt{n}) \); choose + for \( H^+ \) and choose - for \( H^- \) (Extra Credit: prove it); \( c \) is referred to as the \( t \)-statistic.

Two-sided \( z \)-test: Suppose \( \mu_0 \) is on the boundary of a two-sided \( z \)-interval having confidence \( q \); the \( p \)-value for the null hypothesis \( \mu = \mu_0 \) is defined as \( p = 1 - q \) (\( H_0 \) places \( \mu \) outside the \( z \)-interval). It follows (see confidence intervals) that
\[
p = P(|\eta| \geq c) = 2 \frac{1}{\sqrt{2\pi}} \int_{c}^{\infty} e^{-\frac{x^2}{2}} dx
\]
where \( c = (\pi - \mu_0)/\sigma(\sqrt{n}) \) (Extra Credit: prove it); \( c \) is referred to as the \( z \)-statistic.

One-sided \( z \)-test: Suppose \( \mu_0 \) is on the boundary of a one-sided \( z \)-interval having confidence \( p \); the \( p \)-value for the null hypothesis (either \( H^+ \) or \( H^- \), whichever places \( \mu \) in the \( z \)-interval) is defined as \( p \). It follows (see confidence intervals) that
\[
p = P(\eta \geq c) = \frac{1}{\sqrt{2\pi}} \int_{c}^{\infty} e^{-\frac{x^2}{2}} dx
\]
where \( c = \pm(\pi - \mu_0)/(\sigma/\sqrt{n}) \); choose + for \( H^+ \) and choose - for \( H^- \) (Extra Credit: prove it); \( c \) is referred to as the \( z \)-statistic.

The plausibility of \( H_0 \) increases with \( p \)-value; typically \( p < 0.01 \) is taken to mean \( H_0 \) can be rejected, and \( p > 0.1 \) is taken to mean there is insufficient evidence to reject (\( H_0 \) can be accepted). Intermediate \( p \)-values are generally interpreted as indicating an inconclusive test.

Hypothesis tests are often defined with respect to a \textit{significance level} \( \alpha \); the null hypothesis is accepted/rejected according as to whether the \( p \)-value is larger/smaller than \( \alpha \). The significance level is also referred to as the probability of a \textit{Type I error}. A \textit{Type I error} occurs when \( H_0 \) is rejected when it is true. A \textit{Type II error} occurs when \( H_0 \) is accepted when it is false.

Examples (Data Set 1)

9. Is \( H_0: \mu = 13 \) accepted with significance \( \alpha = 0.10 \)? What is the \( p \)-value?

Using \( n = 12, \pi = 13.25, s^2 = 0.33727, \) and \( \mu_0 = 13, \) the \( t \)-statistic and \( p \)-value are
\[
c = \frac{|\pi - \mu_0|}{(s/\sqrt{n})} \approx 1.49122
\]
\[
p = 2 \frac{\Gamma(\frac{12}{2})}{\sqrt{\pi(12-1)}\Gamma(\frac{12-1}{2})} \int_{c}^{\infty} \left(1 + \frac{x^2}{12-1}\right)^{-12/2} dx \approx 0.164
\]

Since \( p > 0.10 \), the null hypothesis is accepted.
10. For what values of $\mu_0$ is $H_0 : \mu < \mu_0$ rejected with significance $\alpha = 0.01$?

Solving

$$0.01 > p = \frac{\Gamma\left(\frac{12}{2}\right)}{\sqrt{\pi(12 - 1)}} \int_{c'} \left(1 + \frac{x^2}{12 - 1}\right)^{-12/2} dx$$

gives $c \geq 2.71808$. Using $n = 12$, $\bar{x} = 13.25$, and $s^2 = 0.33727$, solving

$$2.71808 \leq (\bar{x} - \mu_0)/(s/\sqrt{n})$$

gives $\mu_0 \leq 12.7943$.

11. Assuming $\sigma = .58$, for what values of $\mu_0$ is $H_0 : \mu > \mu_0$ accepted with significance $\alpha = 0.01$?

Solving

$$0.01 < p = \frac{1}{\sqrt{2\pi}} \int_{c'} e^{-\frac{x^2}{2}} dx$$

gives $c \leq 2.32635$. Using $n = 12$, $\bar{x} = 13.25$, and $\sigma = .58$, solving

$$2.32635 \geq - (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$$

gives $\mu_0 \leq 13.6395$.

**Homework**

- Repeat all examples, using Data Set 2.
- For examples 1 and 3 (using Data Set 2), what is the standard error?