Mathematics 50
Probability and Statistical Inference
Winter 1999

Principles of Statistical Estimation

Week 7
February 15-19

1. Parameter estimation

In a typical probabilistic problem parameters of the underlying distribution are given. For example, in one of the first probabilistic problem of Robber and Police the rate of the exponential distribution, λ which was used to model the police time arrival after the robber breaks in, is given. In real life nobody gives you that value. You should derive the parameter value from observations, data. For instance, in police file you could discover the following data:

Break in NYC, Bronx	**Time of police arrival after alarm starts, Min**
Sep 1	3.4
Sep 23	2.4
Oct 6	4.2
Oct 18	10.2
Oct 30	7.8
Nov 7	3.1
Dec 13	5.2
Dec 28	7.2
Jan 3	3.9
Jan 30	2.9
Feb 2	9.5

The question is: what is λ?

The problem of parameter (or point) estimation: We have data $x_1, x_2, ..., x_n$. The number of observations n is called *sample size*. It is assumed that x_i are iid. In other words, observations x_i are independently drawn from the *same* general population with certain distribution function. This distribution function is not known completely, i.e. the family is known but parameters are not. For example, it may be known (assumed) that the distribution is exponential (family of exponential distribution functions), but λ is unknown. How to estimate an unknown parameter λ?

It is typical is statistical theory to denote the unknown parameters as θ.

The estimate must be a function of the data, i.e. $x_1, x_2, ..., x_n$, i.e.

$$\widehat{\theta} = \widehat{\theta}(x_1, ..., x_n).$$

A function of the data (observations) is called *estimator* (or *statistic*). A specific value of the estimator is called *estimate*.

Estimator is a random variable but estimate is not, it is a fixed number.

Consequently, an estimator has a distribution, a mean and a variance.

What is a good estimator?

We do not know the true parameter θ. The only we know that θ belongs to a *parameter space*, e.g. θ is positive. A good estimator $\widehat{\theta}$ must be close to θ for any possible data. But data is random. It implies that we have to consider average (expected values) to characterize the quality of estimators.

Definition 1.1. *An estimator $\widehat{\theta} = \widehat{\theta}(x_1, ..., x_n)$ is called unbiased if*

$$E\left(\widehat{\theta}\right) = \theta.$$

The average of an unbiased estimator is the true parameter.

Why the concept of unbiased estimate does not make sense?

The bias is

$$E\left(\widehat{\theta}\right) - \theta.$$

We can speak of positive or negative bias.

How close is an unbiased estimator to θ?

The quality of an unbiased estimation is measured by

$$var(\widehat{\theta}).$$

The quality of any estimator is measured by Mean Square Error (MSE):

$$MSE = E(\widehat{\theta} - \theta)^2.$$

MSE=Var for an unbiased estimator.

MSE combines variance and bias:

$$MSE = var + (E(\widehat{\theta}) - \theta)^2.$$

Definition 1.2. *An estimator is efficient if*

$$var(\widehat{\theta}) = \min.$$

An estimator is a function of data. What is the simplest function of $x_1, ..., x_n$? Linear function.

Definition 1.3. *An estimator is called linear if it is a linear function of $x_1, ..., x_n$, i.e.*

$$\widehat{\theta} = \lambda_1 x_1 + ... + \lambda_n x_n = \sum_{i=1}^{n} \lambda_i x_i$$

where $\lambda_1, ..., \lambda_n$ are fixed.

Now we apply our theory to estimation of the mean.

2. Estimation of the mean, Rice 10.4.1

Let $x_1, x_2, ..., x_n$ are iid, i.e. drawn independently from a general population with unknown mean μ, *population mean* (or arithmetic mean). What might be a reasonable estimator for μ? The average, or *sample mean (or first sample moment)*:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The average is an unbiased estimator of μ :

$$\begin{aligned} E(\overline{x}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n}E\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}n\mu \\ &= \mu. \end{aligned}$$

Theorem 2.1. *The average is an efficient estimator of μ within the class of unbiased linear estimator of μ.*

Proof. Let $\widehat{\theta}$ be a linear estimator, i.e.

$$\widehat{\theta} = \lambda_1 x_1 + ... + \lambda_n x_n.$$

The unbiasedeness means that

$$\begin{aligned} E(\widehat{\theta}) &= E\left(\sum_{i=1}^{n} \lambda_i x_i\right) = \sum_{i=1}^{n} \lambda_i E(x_i) = \sum_{i=1}^{n} \lambda_i \mu = \mu \sum_{i=1}^{n} \lambda_i \\ &= \mu. \end{aligned}$$

This implies

$$\sum_{i=1}^{n} \lambda_i = 1.$$

Find the variance

$$\begin{aligned} var(\widehat{\theta}) &= var\left(\sum_{i=1}^{n} \lambda_i x_i\right) = \sum_{i=1}^{n} var(\lambda_i x_i) = \sum_{i=1}^{n} \lambda_i^2 var(x_i) = \sum_{i=1}^{n} \lambda_i^2 var(x_i) \\ &= \sigma^2 \sum_{i=1}^{n} \lambda_i^2 \end{aligned}$$

Calculus problem: find $\lambda_1, \lambda_2, ..., \lambda_n$ such that $\sum_{i=1}^{n} \lambda_i = 1$ and

$$\sum_{i=1}^{n} \lambda_i^2 = \min.$$

The maximum is when

$$\lambda_i = const = \frac{1}{n}.$$

Illustrate geometrically.

This leads to the average

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Theorem 2.2. *It can be proven that $\overline{x}$ is efficient within the class of ALL unbiased estimators if x_i are drawn from a normal population.*

Summary. Sample mean is the best among simplest (linear) estimators of the population mean. Also, it is the best if the general population is normal. Otherwise it may not.

Example. Compare the "primitive"estimator of the mean,

$$\frac{1}{2}(x_1 + x_n)$$

with the sample mean.

Primitive estimator is unbiased as well because

$$E\left(\frac{1}{2}(x_1 + x_n)\right) = \frac{1}{2}E(x_1 + x_n) = \frac{1}{2}E(\mu + \mu) = \mu.$$

But it is not so efficient as $\overline{x}$ is:

$$var\left(\frac{1}{2}(x_1 + x_n)\right) = \frac{1}{4}(var(x_1) + var(x_n)) = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{1}{2}\sigma^2.$$

because

$$var(\overline{x}) = \frac{1}{n}\sigma^2 < \frac{1}{2}\sigma^2$$

if $n > 2$.

Definition 2.3. *The SD of the average is called Standard Error,*

$$SE = \frac{\sigma^2}{n}.$$

Standard error <SD of a single observation (less by square root of n), that is why we use average in statistics.

Mean is a characteristic of location. You know, there are other parameters of location. Name them

3. Median as an alternative to average, Rice 10.4.2

Sample mean (average) is the best if observations are drawn from a normal population. What if not? For example, what happens to the sample mean if there is an outlier, i.e. unusual observation? Consider the following example. Below is reported monthly income of 9 people (in thousands of dollars)

$$3.2, 2.9, 1.7, 3.4, 4.1, 2.6, 0, 3.9, 2.5$$

We want to estimate the population monthly income. We have

$$\frac{1}{9}(3.2 + 2.9 + 1.7 + 3.4 + 4.1 + 2.6 + 0 + 3.9 + 2.5) = 2.7$$

Can we say that the population monthly income is about \$2, 700? What is wrong? Zero is under question (that guy is unemployed). We may exclude zero

$$\frac{1}{8}(3.2 + 2.9 + 1.7 + 3.4 + 4.1 + 2.6 + 3.9 + 2.5) = 3.04$$

It is called *trimmed* mean.

But what if we get not zero but 0.1, 0.4. Should we every time decide to include of exclude?

There is an intelligent way to provide a robust estimation of location parameter: median.

Median is such a value that the number of observations at the left is equal to the number of observation at the right. To find the median we have to order our observations, then to find the $n/2$th observation from the left:

$$\begin{aligned} \widetilde{x} &= median(3.2, 2.9, 1.7, 3.4, 4.1, 2.6, 0, 3.9, 2.5) \\ &= median(0, 1.7, 2.5, 2.6, 2.9, 3.2, 3.4, 3.9, 4.1) \\ &= 2.9 \end{aligned}$$

Why median is a robust estimator? Let we have $x_1, ...x_n$ and x_n is an outlier, i.e.e x_n is big.

Sample mean:

$$\overline{x} = \frac{1}{n}(x_1 + ...x_{n-1} + x_n) \to \infty \text{ if } x_n \to \infty.$$

Median is robust to outliers:

$$\widetilde{x} = const \text{ if } x_n \to \infty$$

because it only counts how many values are at the right.

Median is a *robust* estimator of the population center because it is not affected by outliers.

There are other robust estimators, they are called M-estimators (Rice 10.4.4).

4. My first confidence interval

$\overline{x}$ estimates μ. It is a point estimation. If $\overline{x} = 1.56$ can we claim that $\mu = 1.56$? No, because $\overline{x}$ is a random variable and $\overline{x}$ may take value 1.78 or 1.31 using other data. In fact, $\Pr(X = 1.56) = 0$ if X is continuous.

So, how to make statistical inference about μ? What the values of μ might be?

Confidence intervals are used to provide an *interval estimation* of population parameters. The idea is to find such an interval, based on the point estimation, that it covers an unknown population parameters with certain (predefined) probability.

Remember, data is random, so our statistical inference has certain probability.. We cannot find an interval that covers an unknown parameter with probability 1. That is why we should start with specifying the *coverage probability*, λ. It is common to take $\lambda = .95$, i.e. we speak of 95% confidence interval. Sometimes λ is called *confidence level*.

$1 - \lambda = \alpha$ is called *significance level*.

Significance Level = 1 - Confidence Level.

Let us assume that $x_1, x_2, ..., x_n$ are iid from a normal distribution with **known** variance σ^2, i.e.

$$x_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, ..., n$$

where μ is unknown and σ^2 is known.

Examples?

We know that

$$\overline{x} \sim \mathcal{N}(\mu, \sigma^2/n)$$

Also, we remember that if $y \sim \mathcal{N}(mean, SD^2)$

$$\Pr(|y - mean| < 1.96 \times SD) = .95$$

Rewriting this as

$$\Pr(|\overline{x} - \mu| < 1.96 \times \frac{\sigma}{\sqrt{n}} = .95$$

we obtain the **95% CI** for μ :

$$\Pr\left(\overline{x} - 1.96 \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right) = .95$$

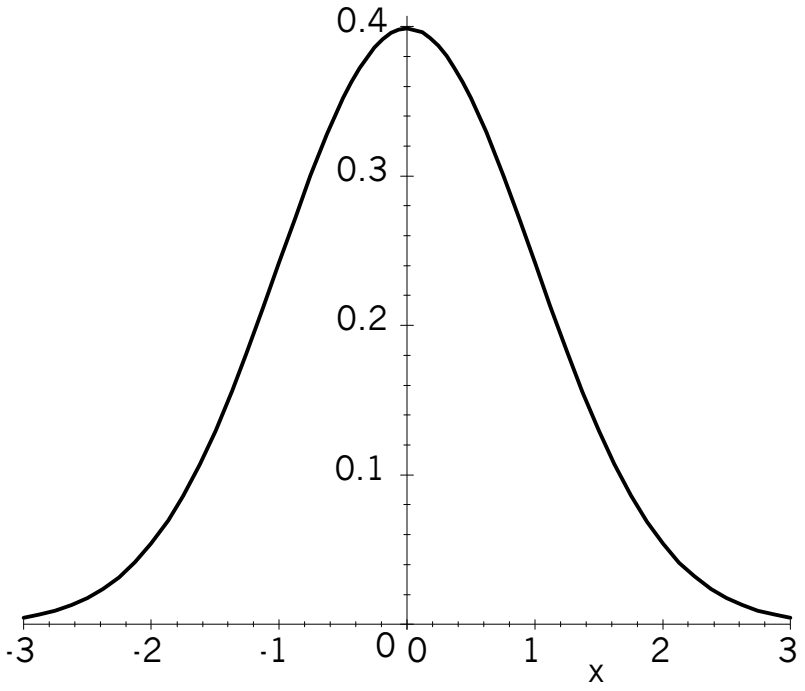
CI is

$$\left(\overline{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right).$$

Interpretation of 95% CI: if you have 1000 trials of n observations $x_1, x_2, .., x_n$ and you calculate CI then it will cover the true μ in 950 cases.

Illustrate geometrically.

How to find CI for any given confidence level, $\lambda = 1 - \alpha$?



Let $x_\alpha > 0$ such that $\Phi(|X| > x_\alpha) = \alpha$.
It means that $p = 1 - \alpha/2$ and $\Pr(X < x_\alpha) = p$.
The the (1-α)100% CI is

$$\left(\overline{x} - x_\alpha \times \frac{\sigma}{\sqrt{n}}, \overline{x} + x_\alpha \times \frac{\sigma}{\sqrt{n}}\right)$$

Find 90% CI:
$p = 1 - .1/2 = .95$
$x_{.1} = 1.65$ and

$$\left(\overline{x} - 1.65 \times \frac{\sigma}{\sqrt{n}}, \overline{x} + 1.65 \times \frac{\sigma}{\sqrt{n}}\right).$$

5. Estimation of the variance, Rice 7.32

Let $x_1, x_2, ..., x_n$ are iid, i.e. drawn independently from a general population with unknown mean μ and variance σ^2 (population variance). What might be a reasonable estimator for σ^2? The sample variance (or second sample centered moment):

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Is it unbiased? No.

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad = \quad \sum_{i=1}^{n}(x_i^2 - 2\overline{x}x_i + \overline{x}^2) = \sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n\overline{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + n\overline{x}^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

so that

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2.$$

Continue,

$$\begin{aligned} E\left(\widehat{\sigma}^2\right) &= \frac{1}{n}E\left(\sum_{i=1}^{n} x_i^2 - \overline{x}^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} x_i^2\right) - E\left(\overline{x}^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left(x_i^2\right) - E\left(\overline{x}^2\right). \end{aligned}$$

But

$$E\left(x_i^2\right) = \sigma^2 + \mu^2$$

and

$$\begin{aligned} E\left(\overline{x}^2\right) &= \frac{1}{n^2}E\left(\sum x_i\right)^2 = \frac{1}{n^2}E\left(\sum_i \sum_j x_i x_j\right) \\ &= \frac{1}{n^2}E\left(\sum_{i\neq j} x_i x_j + \sum_i x_i^2\right) \\ &= \frac{1}{n^2}\left(\sum_{i\neq j} Ex_i Ex_j + \sum_i Ex_i^2\right) \\ &= \frac{1}{n^2}\left(\sum_{i\neq j} \mu^2 + \sum_i (\mu^2 + \sigma^2)\right) = \frac{1}{n^2}\left(n^2\mu^2 + n\sigma^2\right) \\ &= \mu^2 + \frac{1}{n}\sigma^2. \end{aligned}$$

Finally,

$$E\left(\widehat{\sigma}^2\right) = (\sigma^2 + \mu^2) - (\mu^2 + \frac{1}{n}\sigma^2) = (1 - \frac{1}{n})\sigma^2.$$

This means that $\widehat{\sigma}^2$ is negative biased. It will be unbiased if

$$\frac{n\widehat{\sigma}^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

The **unbiased** version of sample variance:

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Estimation of SD

$$\widehat{\sigma} = \sqrt{\widehat{\sigma}^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

When the biased and unbiased version of σ^2 are close: when n is large.

Large vs. small sample size - common issue in statistics.

6. Confidence interval for the mean in large samples

Come back to our problem of CI construction for the mean.

$x_1, x_2, ..., x_n$ are iid with unknown population mean and unknown variance σ^2.

From CLT we know that

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

But we can estimate $\widehat{\sigma}$ as

$$\widehat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

so the 95% CI for μ is approximately

$$\left(\overline{x} - 1.96 \times \frac{\widehat{\sigma}}{\sqrt{n}}, \overline{x} + 1.96 \times \frac{\widehat{\sigma}}{\sqrt{n}}\right).$$

It works only for large samples (quick and dirty CI). Why?

Because we used an estimate instead of the true value of σ. In small samples theory must change (more complicated).

Example (continued).

$\frac{1}{9}(3.2 + 2.9 + 1.7 + 3.4 + 4.1 + 2.6 + 0 + 3.9 + 2.5) = 2.7$

$\frac{1}{9}((3.2 - 2.7)^2 + (2.9 - 2.7)^2 + (1.7 - 2.7)^2 + (3.4 - 2.7)^2 + (4.1 - 2.7)^2$

$+(2.6 - 2.7)^2 + (0 - 2.7)^2 + (3.9 - 2.7)^2 + (2.5 - 2.7)^2) = 1.3911$

i.e.

$$\begin{aligned} \widehat{\sigma}^2 &= 1.4 \\ \widehat{\sigma} &= 1.18 \end{aligned}$$

With 95% probability monthly income in US is within the interval

$$\begin{aligned} & \left(\overline{x} - 1.96 \times \frac{\widehat{\sigma}}{\sqrt{n}}, \overline{x} + 1.96 \times \frac{\widehat{\sigma}}{\sqrt{n}}\right) \\ = & \left(2.7 - 1.96 \times \frac{1.18}{\sqrt{9}}, 2.7 + 1.96 \times \frac{1.18}{\sqrt{9}}\right) \\ = & (1.93, 3.47). \end{aligned}$$

Interpretation: take 1000 people at random from US population, ask their monthly income. Then you can expect that 950 (approx..) incomes fall within $(1.93, 3.47)$.

Statistics: how by analysis of a (relatively) small sample we can draw inferences on general population.

How to make inference more precise, i.e. narrow CI?

Hawaii problem. Fred works as real estate agent for 7 years. His history earnings is:

Year	Earnings, thousands
1992	23
1993	16
1994	37
1995	28
1996	31
1997	29
1998	43

He plans to go on vacation to Hawaii next year. He estimates it requires minimum \$3,000 to go to Hawaii. He can afford not more then 7% of his annual income. Evaluate the probability he will go to Hawaii next year.

Solution. Let X be his earnings next year, thousands dollars. The the asked probability is

$$\Pr(.07X_{1999} > 3)$$

or

$$\Pr(X_{1999} > 42.86).$$

We estimate his annual income as the average

$$\overline{x} = \frac{1}{7}(23 + 16 + 37 + 28 + 31 + 29 + 43) = 29.57$$

We estimate his annual variance as

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{7}((23 - 29.57)^2 + (16 - 29.57)^2 + (37 - 29.57)^2 + (28 - 29.57)^2 \\ &\quad +(31 - 29.57)^2 + (29 - 29.57)^2 + (43 - 29.57)^2) \\ &= 66.8 \end{aligned}$$

and $\widehat{\sigma} = \sqrt{66.8} = 8.17$.

Thus, roughly we can approximate distribution of X_{1999} as

$$X_{1999} \sim N(29.57, 8.17).$$

So that the asked probability translates into

$$\begin{aligned} \Pr(X_{1999} &> 42.86) = 1 - \Pr(X_{1999} \leq 42.86) = 1 - \Phi\left(\frac{42.86 - 29.57}{8.17}\right) \\ &= 1 - \Phi(1.62) = 1 - .95 = .05 \end{aligned}$$

Chances are .5 out of 100.

7. Estimation of distribution and density functions

Let we have

$$x_1, x_2, ..., x_n \text{ are iid}$$

They came from the same distribution function and density. Can we restore (estimate) theses functions.

7.1. Empirical distribution function (Rice 10.2.1)

Recall, d.f. is defined for each x as

$$F(x) = \Pr(x_i \leq x) \simeq \frac{\#\ (x_i \leq x)}{\#\ \text{events}} = \frac{1}{n}(\#x_i \leq x).$$

The **algorithm** to estimate d.f. F is as follows:

1. Order observations.
2. Grid y-axis with the step length 1/n.
3. Scan the order observations and make a jump.

Where is median?

Given p, what is quantile x_p?

7.2. Histogram and density estimation

How to estimate the density?

Recall

$$f(x) = F'(x) = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} \simeq \frac{1}{n\Delta x} \times \#(x \leq x_i \leq x + \Delta x)$$

Histogram is proportional to # of observations within interval

if the length of intervals (classes, bins) are the same.

Algorithm to construct histogram:

1. Order observations.
2. Divide the range, $\min x_i, \max x_i$ into n intervals.
3. Find # observations which fell into an appropriate interval (histogram frequency).

Density estimation: one can assume that the density is a mixture of normal densities, so that we can apply the idea of running interval (kernel estimation). The width of that interval/window is called bandwidth.

7.3. Q-Q plot

One of the major assumption in statistics is that the distribution is normal.

How to check this?

Quantile-Quantile (Q-Q) plots are for this purpose. Q-Q plot is the plot of empirical Quantiles vs. theoretical Quantiles derived from normal distribution. If $x_1, ..., x_n$ are iid and drawn from a normal population then the according curve must be close to the 45^o line.

How do we make Q-Q plot? First, compute

$$\begin{aligned} \widehat{\mu} &= \overline{x}, \\ \widehat{\sigma} &= \sqrt{\frac{1}{n}\sum (x_i - \overline{x})^2}. \end{aligned}$$

If x are from normal population then

$$x_i \simeq N(\overline{x}, \widehat{\sigma}^2).$$

and

$$z_i = \frac{x_i - \widehat{\mu}}{\widehat{\sigma}} \simeq N(0,1).$$

How to make a Q-Q plot:

We take probabilities 1/n, 2/n, 3/n,..., (n-1)/n. The according quantiles for the empirical distribution are $x_1, x_2, ..., x_{n-1}$.

The Quantiles for the normal distribution are

$$\widehat{\sigma} z_1 + \widehat{\mu}, \widehat{\sigma} z_2 + \widehat{\mu}, ..., \widehat{\sigma} z_n + \widehat{\mu}$$

where:

z_1 is the quantile of standard normal distribution $\Phi(z_1) = 1/n$,

z_2 is the quantile of standard normal distribution $\Phi(z_2) = 2/n$,

...

z_{n-1} is the quantile of standard normal distribution $\Phi(z_{n-1}) = (n-1)/n$.

If observations came form a normal distribution then we should see that the points are close to 45% line.

Quantile-quantile plot

#	x_i	z_i, Quantiles of $N(0,1)$	$\widehat{\sigma} z_i + \widehat{\mu}$ based on $N(\widehat{\mu}, \widehat{\sigma}^2)$
1	16	-1.0675705	20.14335
2	23	-0.5659488	24.57267
3	28	-0.1800124	27.98049
4	29	-0.1800124	31.15951
5	31	0.1800124	34.56733
6	37	0.5659488	38.99665

8. Homework (due Feb 24)

1. (4 points). Prove that Mean Square Error is equal to the sum of variance and squared bias.

Solution. We need to prove that $E(\widehat{\theta} - \theta)^2 = var(\widehat{\theta}) + (E(\widehat{\theta}) - \theta)^2$ for any estimator $\widehat{\theta}$ and any fixed θ. Denote $E(\widehat{\theta}) = \mu$. Then we have

$$\begin{aligned} E(\widehat{\theta} - \theta)^2 &= E((\widehat{\theta} - \mu) + (\mu - \theta))^2 = E\left((\widehat{\theta} - \mu)^2\right) + 2\left(E(\mu - \theta)(\widehat{\theta} - \mu)\right) + E\left((\mu - \theta)^2\right) \\ &= var(\widehat{\theta} - \mu)^2 + 2(\mu - \theta)E(\widehat{\theta} - \mu) + (\mu - \theta)^2. \end{aligned}$$

But $E(\widehat{\theta} - \mu) = E(\widehat{\theta}) - \mu = \mu - \mu = 0$, so we come the needed equality.

2. (6 points). There are two independent uniformly distributed RVs X_1 and X_2 on $(0, a)$. Is $\max(X_1, X_2)$ an unbiased estimator of a? What is an unbiased estimator of a?

Solution. The df for the maximum is $F^2(x)$ where $F(x)$ is df of $U(0,a)$. But $F(x) = a^{-1}x$ where $0 < x < a$. Thus, we obtain the df of $Y = \max(X_1, X_2)$ is $a^{-2}x^2$ and the density is $2a^{-2}x$. The expectation of Y is

$$E(Y) = 2a^{-2}\int_0^a xxdx = 2a^{-2}\int_0^a x^2dx = \frac{2}{3}a^{-2}a^3 = \frac{2}{3}a.$$

Thus, $E(Y)$ is less than a, so that Y is a *biased* estimator of a. To make it unbiased we should take $1.5\max(X_1, X_2)$.

3. (5 points). Are mean and median equivariant to linear transformations? In other words, if data are linearly transformed as $a + bX$ do mean and median transform in the same way?

Solution. Let the data are given as $x_1, x_2, ..., x_n$ with the sample mean $\overline{x}$ and median m. Linear transformation on data leads to data as $a + bx_i$. The sample mean of the transformed data is

$$\frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = \frac{1}{n}(na + b\sum_{i=1}^{n}x_i) = a + b\overline{x},$$

so that the mean transforms in the same way. Now let us prove the same for median. Let us first assume $b > 0$. Median m is such that $n/2$ observations have values less than m and $n/2$ observations have values more than m. If we multiply x_i by b the median multiplies by b as well because the order of x_i remains unchanged. Also if we add any number a to data nothing changes it terms of the order. If $b < 0$ the order flips but the same $n/2$ observations will be at left and $n/2$ will be at right. So, the median transforms in the same way as $ax_i + b$.

4. (6 points).The grades for one of the homework are: 38, 31.5, 25.5, 34, 37, 36.5, 32.5, 36.5, 33, 27, 37, 35.5, 31.5, 36.5, 13.5, 36.5, 34, 16, 33, 31.5, 33.5 (maximum # points = 38). Compute mean, median and standard deviation. Should we excluded some grades to make inference more robust?

Solution. Mean=31.9, Median=33.5, SD=6.55. An obvious candidates for outlier got 13.5 and 16. There is no rule to decide whether they should be removed from the calculations. I would leave them because they are part of the class and one could expect similar performance of those guys in future.

5. (5 points). What is an approximate 95% CI for the average grade for the next homework assuming the average remains the same and grades follow normal distribution.

Solution. The SD for mean is $\widehat{\sigma}/\sqrt{n}$ and 95%CI is

$$\left(\overline{x} - 1.96\frac{\widehat{\sigma}}{\sqrt{n}}, \overline{x} + 1.96\frac{\widehat{\sigma}}{\sqrt{n}}\right)$$

For our data $\widehat{\sigma} = 6.55$ and $n = 21$. Thus 95% CI for the mean is (29.1, 34.7).

6. (6 points). Evaluate the number of students to get more than 90% in the next homework using normal approximation. List assumptions you made to come up with this number. Is there any way to project this number without normal assumption?

Solution. We assume that $x_i \sim N(31.9, 6.55^2)$. 90% grade corresponds to .9*38=34.2 points. The probability to get more than 34.2 points is $\Pr(X > 34.2)$ where $X \sim N(31.9, 6.55^2)$. It is calculated as follows

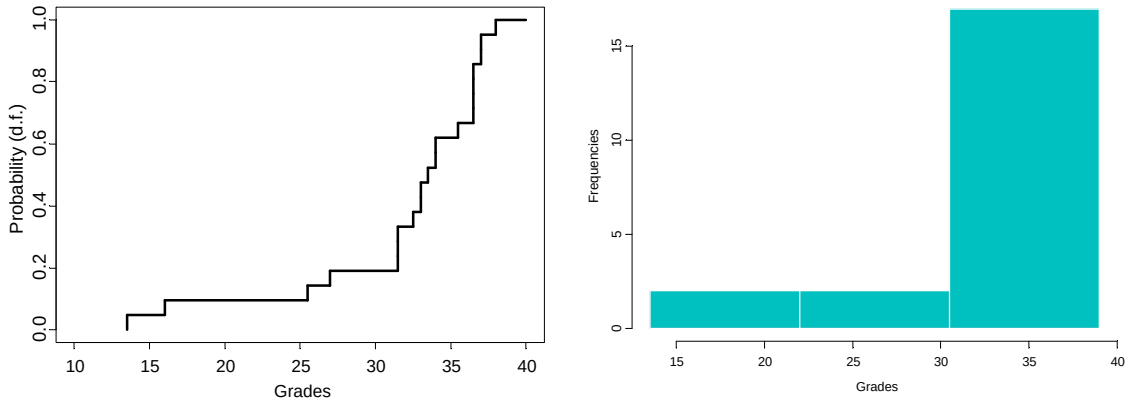
$$1 - \Pr(X < 34.2) = 1 - \Phi\left(\frac{34.2 - 31.9}{6.55}\right) = 1 - \Phi\left(.35115\right) = .36.$$

Since $n = 21$ the expected number of students to get more than 90% is $21 \times .36 \simeq 8$ students. As follows from empirical distribution function this # is around 8 as well.

7. (7 points). Plot empirical distribution function for data from 4. What is 50% CI based on the lower and upper quartiles. What is 50% CI based on the normal distribution assumption.

Solution. $x_{.25} = 31.5$ and $x_{.75} = 36.5$. Using normal assumption we obtain the following 50% CI

$$(\overline{x} - .67\widehat{\sigma}, \overline{x} + .67\widehat{\sigma}) = (27.5, 36.3)$$



8. (5 points). Find histogram frequencies and plot the histogram for data from 4 using 3 bars (classes or bins).

Solution. See the above Figure.

9. (7 points). Plot Q-Q graph for data from 4. Do you think the data have normal distribution?

Solution. There are serious doubts that the data came from normal population.