Sampling distribution and parameter estimation

(Week 02 lecture notes)

Wei Q. Deng

Department of Statistical Sciences

July 9th 2018

Recap from last lecture

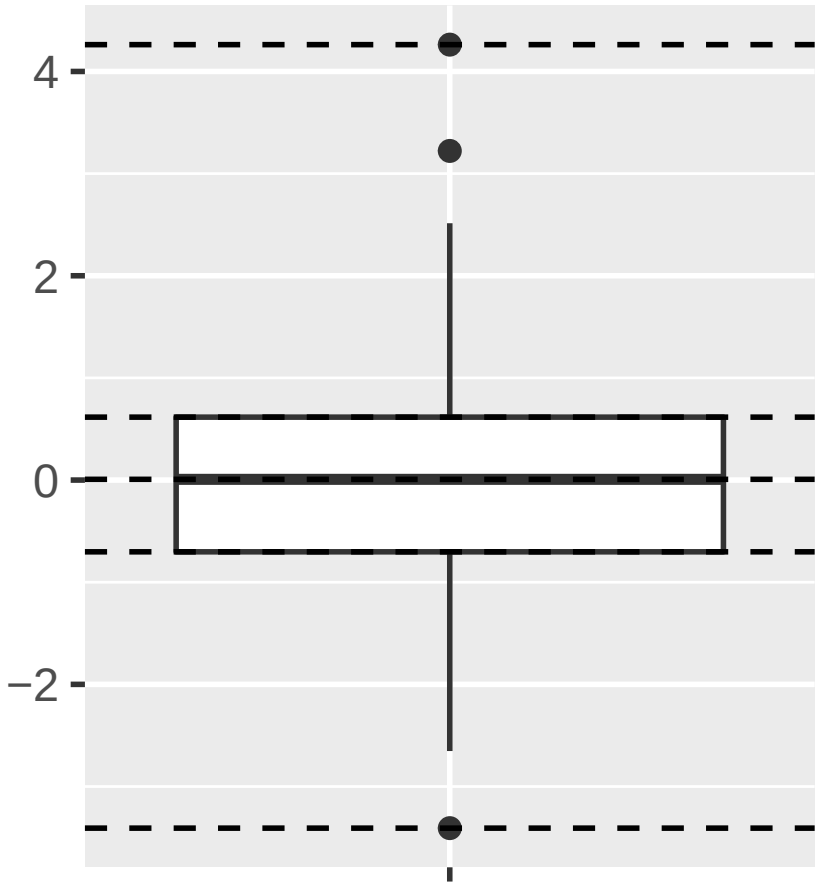
- Descriptive statistics (numerical summaries)
- Boxplot
- Connecting boxplot and histogram (where is the mean/median roughly, what about skewedness?)
- IQR (and other inter quantile range)

Descriptive statistics

- to learn data distribution
- attributes for features of a distribution
  - central tendency
  - dispersion/spread
  - symmetry/skewedness
  - shape of tails
  - modality
  - other anomalies such as missing values and outliers
- a quick way to visualize data distribution with just five numbers
  - maximum
  - 3rd quartile (75% percentile)
  - median (50% percentile)
  - 1st quartile (25% percentile)
  - minimum

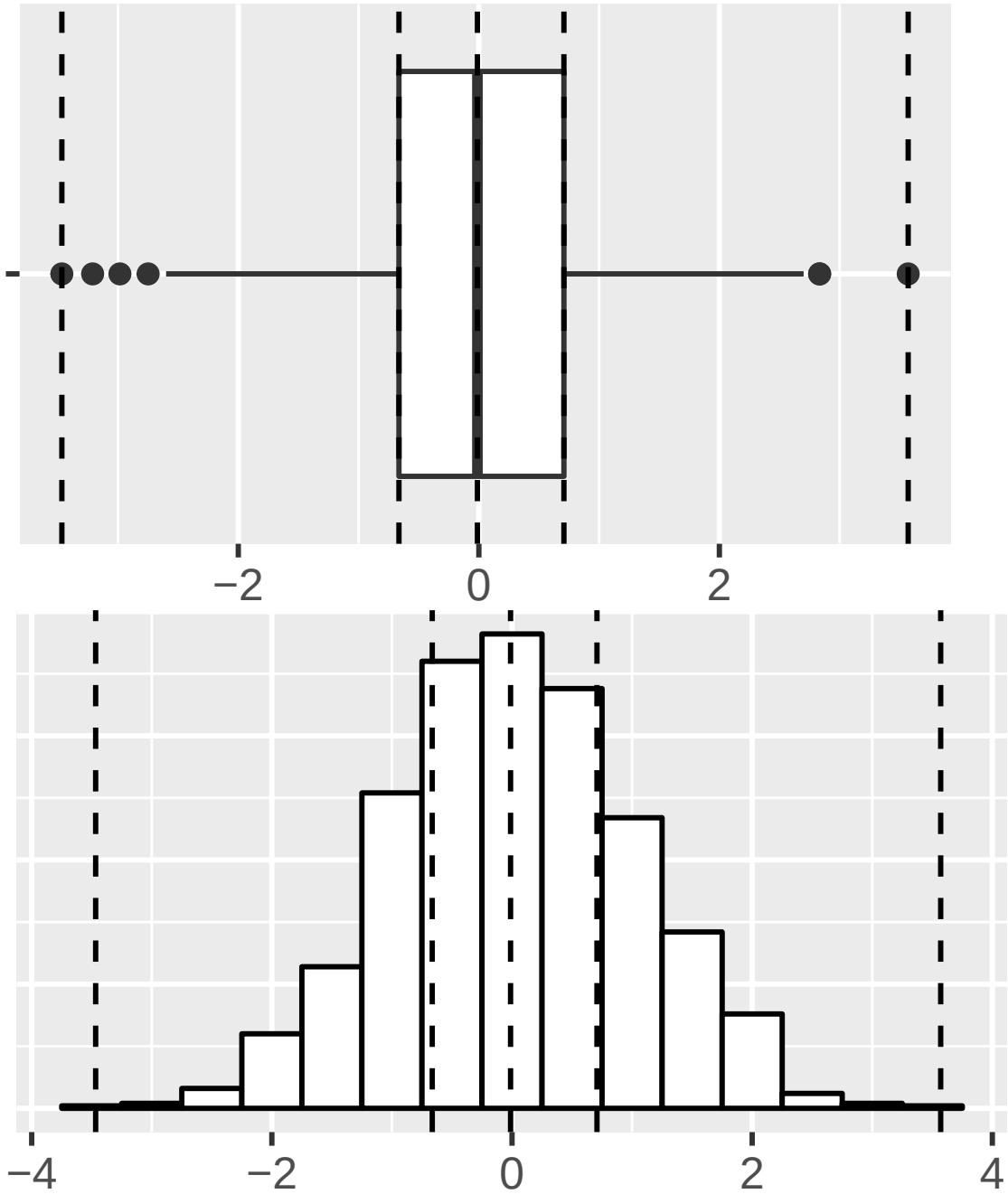
Boxplot

Boxplot provides a visualization of how data behave using the five descriptive statistics.



- the width of the box gives the IQR (50% of the data)
- the length of the whiskers is 1.5 IQR
- all observations outside of the ± 2IQR range (from median) are shown as dots
- outliers are defined by values less than $Q1 - 1.5\text{IQR}$ or greater than $Q3 + 1.5\text{IQR}$

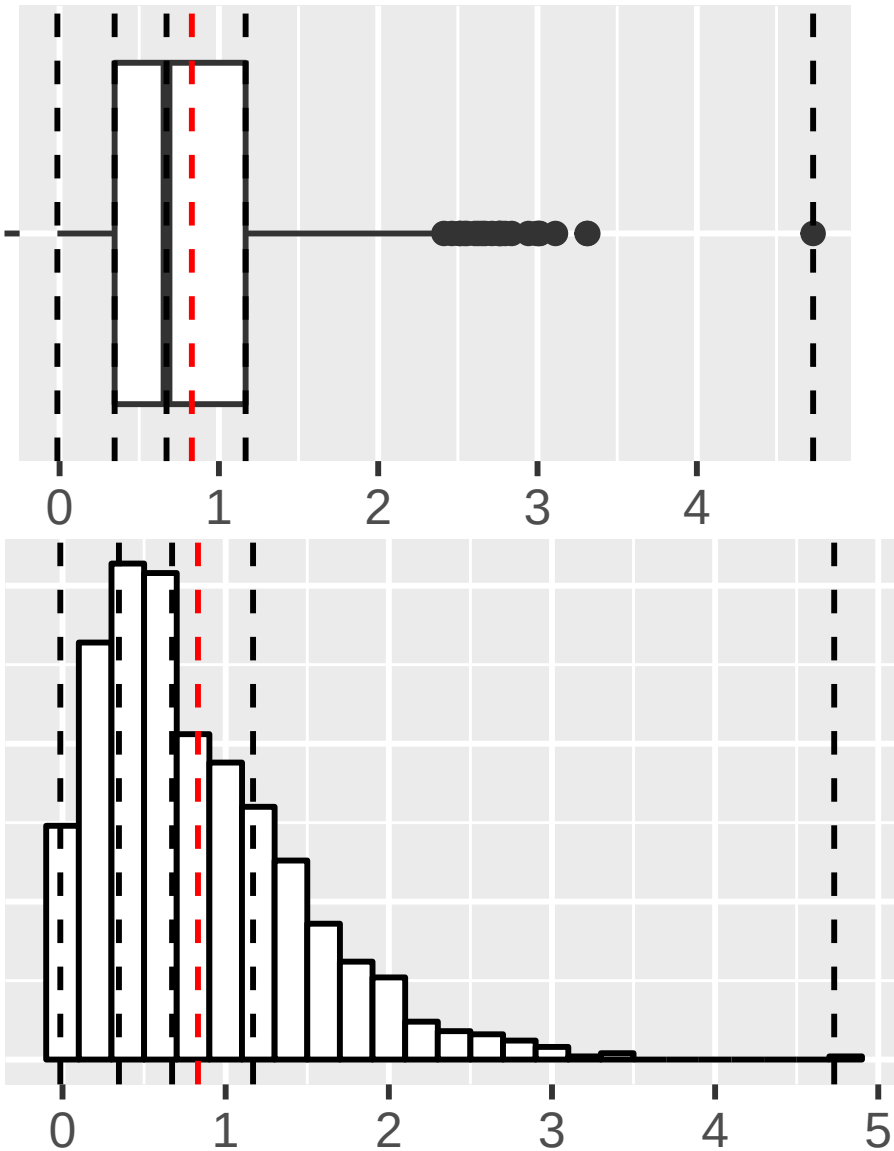
Boxplot and Histogram



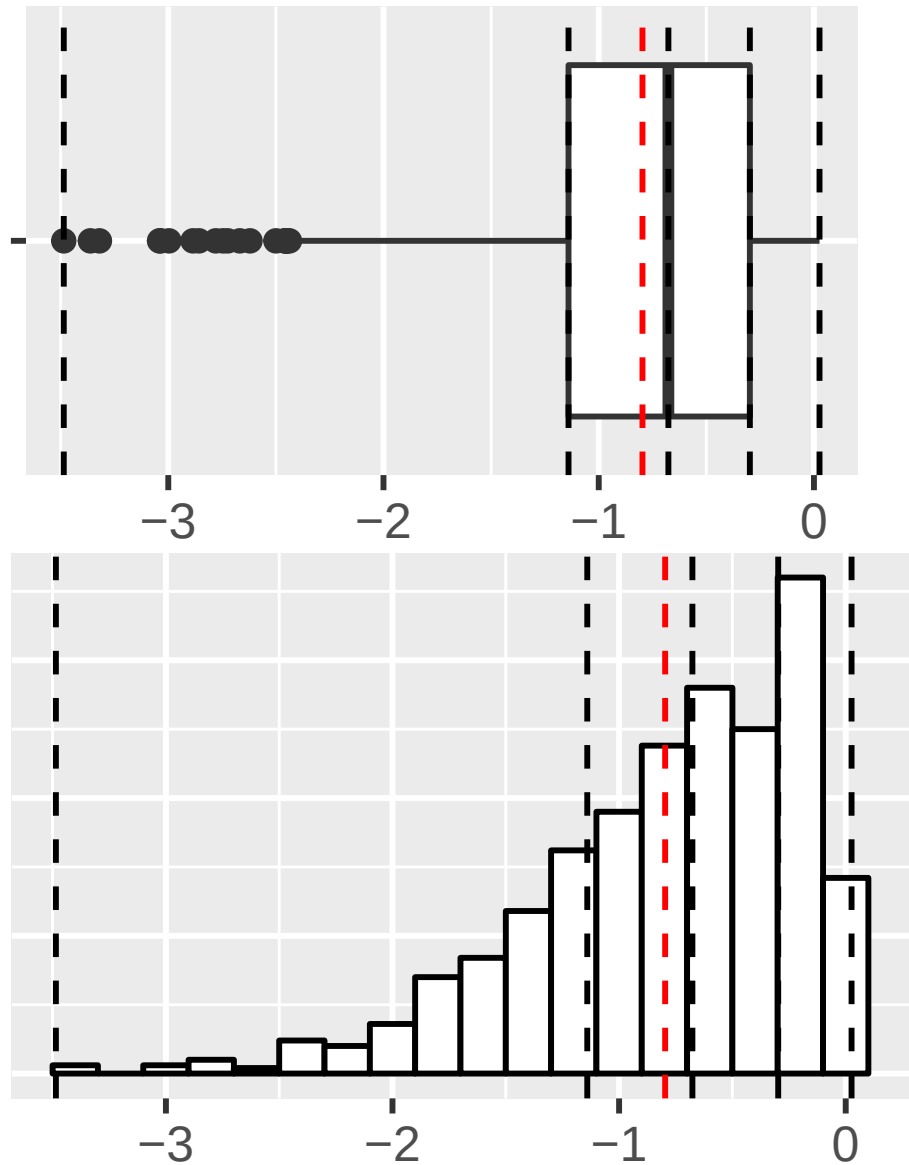
- boxplot gives a better visual for spread and identifying outliers
- histogram shows the overall shape and tail better
- both can be used to identify symmetry/asymmetry

Boxplot and Histogram

Skewed to the right



Skewed to the left



IQR

- Inter-quartile range is a special case of an inter 50% quantile distance. Let

$$x_p = \{x : F_X(x) = p\}$$

and

$$x_{1-p} = \{x : F_X(x) = 1 - p\}$$

for $0 < p < 0.5$

- Then the inter $1 - 2p$ quantile distance is simply $x_{1-p} - x_p$.
- When $p = 0.25$, we recover the IQR $Q_3 - Q_1 = x_{0.75} - x_{0.25}$.

IQR and standard deviation

Since both are measures of spread, they are closely related quantities. But for a normal random variable, the relationship is deterministic due to

$$F_X(x) = \frac{1}{2}\left[1 + \text{err}\left(\frac{x-\mu}{\sigma}\right)\right]$$

- In the case of a symmetric normal distribution with mean centered at $\mu = 0$, $F_X(x)$ reduces to a function of σ alone.
- Note that err function has an analytical form involving integration without a closed-form representation, so the relationship is usually approximated.
- You can then determine the approximated relationship between σ and IQR or more generally inter $1 - 2p$ quantile distance for any arbitrary $0 < p < 0.5$.

Topics covered in this lecture

- Sampling
  - Concept of a random sample
  - Mean of random variables from a random sample
  - Sampling distribution of the mean of random variables
  - Sampling from a normal distribution
- Parameters
  - Finding estimators for a parameter
    - Method of moments estimators
    - Maximum likelihood estimators
  - Evaluating estimators
    - Consistency
    - Unbiasedness
    - Robustness

Concept of a random sample

A formal definition of random sample

A collection of random variables (r.v.) $X_1, \ldots, X_n$ are a *random sample* of size n from the *population* if $X_1, \ldots, X_n$

- are mutually independent
- have the same distribution (either p.d.f or p.m.f)

You can also say a random sample of r.v.s $X_1, \ldots, X_n$ are independent and identically distributed (iid in short) with probability function $f_X(x)$.

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = f_X(x_1) \cdots f_X(x_n) = \Pi_{i=1}^{n} f_X(x_i)$$

what about a random sample from a known parametric family?

- Suppose the random sample comes from a known parametric family with unknown parameter θ (could either be a real value or a vector), e.g.
  - normal distribution $\theta = (\mu, \sigma)$
  - t-distribution $\theta = k$, the degrees of freedom
  - binomial or Bernoulli distribution $\theta = p$, the probability of a positive outcome
- By considering different values of θ, we can observe how a random sample behave for different populations.

$$f_{\mathbf{X}}(x_1, \ldots, x_n|\theta) = \Pi_{i=1}^{n} f_X(x_i|\theta)$$

How to obtain a random sample?

To obtain a truly random sample $(x_1, \ldots, x_n)$, we have to sample from an **infinite** population $(x_1, \ldots, x_N, \ldots)$
For a **finite** population (e.g. $(x_1, \ldots, x_N)$, where $n < N < \infty$), we need to *sample with replacement*

- after each draw, the same value is replaced in the population
- this way, each x_i $(i = 1, \ldots, n)$ has exactly probability $\frac{1}{N}$ of being drawn
- we must have $\mathrm{P}(X_2 = y | X_1 = y) = \frac{1}{N} = \mathrm{P}(X_2 = y)$
- and $\mathrm{P}(X_2 = y' | X_1 = y) = \frac{1}{N}$
- the probability of drawing a sample is $\frac{1}{N^n}$

How to obtain a random sample? (cont'd)

On the other hand, *sampling without replacement* refers to the drawing process

- where after each value is taken, the choice of this particular value becomes unavailable for the next draws.
- we must have $\mathrm{P}(X_2 = y | X_1 = y) = 0 \neq \mathrm{P}(X_2 = y) = \frac{1}{N}$
- and $\mathrm{P}(X_2 = y' | X_1 = y) = \frac{1}{N-1}$ where $(y \neq y')$
- it is clear the probability of drawing a sample is $\frac{1}{N}\frac{1}{N-1}\cdots\frac{1}{N-n+1}$
- a sample resulted from *sampling without replacement* thus does not satisfy the condition of being a random sample (especially if $n < N$ and not $n << N$)

However, notice the marginal distribution $f_{X_i}(x)$ of X_i from *sampling with replacement* (gives a **random sample**) and *sampling without replacement* (does **not** result in a random sample) is the same for $i = 1, \ldots, n$.

An example

Suppose we take a random sample of 10 values from the finite population $\{1, \ldots, 1000\}$ **with replacement**, what is the probability that all sample values are greater than 200?
In this case the samples are mutually independent and we can calculate the exact probability to be:

$$\mathrm{P}(X_1 > 200, \ldots, X_{10} > 200) = \Pi_{i=1}^{10} \mathrm{P}(X_i > 200) = \left(\frac{800}{1000}\right)^{10} \sim 0.107$$

An example (cont'd)

Suppose we take a random sample of 10 values from the finite population $\{1, \ldots, 1000\}$ **without replacement**, what is the probability that all sample values are greater than 200?
In this case the samples are not mutually independent, but this can be translated to a problem of counting the number of samples greater than 200.
Let Y be the number of samples greater than 200, and it has a hypergeometric distribution[1]:

$$\mathrm{P}(X_1 > 200, \ldots, X_{10} > 200) = \mathrm{P}(Y = 10) = \frac{\binom{800}{10}\binom{200}{0}}{\binom{1000}{10}} = 0.106$$

[1] probability of $k = 10$ successes (random draws for which the object drawn has a specified feature, being greater than 200) in $n = 10$ draws, **without replacement**, from a finite population of size $N = 1000$ that contains exactly $K = 800$ objects with that feature, wherein each draw is either a success or a failure.

reflection

- Suppose the finite population is $\{1, \ldots, 100\}$, repeat the calculation. Are the two probabilities different?
- Suppose the finite population remains the same $\{1, \ldots, 1000\}$, but we are taking 100 samples, repeat the calculation. Are the two probabilities different now?
- What do you conclude from these calculations?

Mean of random variables from a random sample

A formal definition of a statistic

Given a sample $X_1, \ldots, X_n$, a well-defined statistic is expressed as a function of the sample $T(X_1, \ldots, X_n)$.

- The function T can be real-valued or a vector
- The statistic itself is a random variable $Y = T(X_1, \ldots, X_n)$
- In some cases, the distribution of Y is tractable

The distribution of Y is derived from the distribution of a sample $X_1, \ldots, X_n$ and thus also called the **sampling distribution** of Y. Generally speaking, the definition of a statistic can be almost anything, but it **cannot** be a function of the (unknown or known) parameter θ.

Sample mean of random variables

Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Define $T(X_1, \ldots, X_n) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $T(X_1, \ldots, X_n) = S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$. Show

1. $\mathrm{E}(\bar{X}) = \mu$
2. $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$
3. $\mathrm{E}(S^2) = \sigma^2$

Try to work these out.

Proof of a)

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

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

$$= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}(X_i) \tag{3}$$

$$= \frac{1}{n} n\mathsf{E}(X_i) \tag{4}$$

$$= \mu \quad \square \tag{5}$$

Proof of b)

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) \tag{6}$$

$$= \frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n} X_i) \tag{7}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \tag{8}$$

$$= \frac{1}{n}\mathrm{Var}(X_i) \tag{9}$$

$$= \frac{\sigma^2}{n} \quad \square \tag{10}$$

Before c), remember the following from STA247

Here we used the following relationship between the mean and variance:

$$\text{Var}(X_i) = \text{E}[(X_i - \mu)^2] = \text{E}[X_i^2 - 2\mu X_i + \mu^2] = \text{E}(X_i^2) - \mu^2$$

similarly

$$\text{Var}(\bar{X}) = \text{E}(\bar{X}^2) - E(\bar{X})^2 = \text{E}(\bar{X}^2) - \mu^2$$

Proof of c)

$$\mathrm{E}(S^2) = \mathrm{E}\Big(\frac{1}{n-1}\Big[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\Big]\Big) \tag{11}$$

$$= \frac{1}{n-1}\mathrm{E}\Big[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\Big] \tag{12}$$

$$= \frac{1}{n-1}\Big[n\mathrm{E}(X_i^2) - n\mathrm{E}(\bar{X}^2)\Big] \tag{13}$$

$$= \frac{n}{n-1}\Big[\mathrm{E}(X_i^2) - \mathrm{E}(\bar{X}^2)\Big] \tag{14}$$

$$= \frac{n}{n-1}\Big[\mathrm{Var}(X_i) + \mathrm{E}(X_i)^2 - \mathrm{Var}(\bar{X}) - E(\bar{X})^2\Big] \tag{15}$$

$$= \frac{n}{n-1}\Big[\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2\Big] \tag{16}$$

$$= \frac{n}{n-1}\Big[\sigma^2 - \frac{\sigma^2}{n}\Big] \tag{17}$$

$$= \sigma^2 \quad \square \tag{18}$$

Sampling distribution of the mean of random variables

Sampling distribution of the mean of random variables

Let $X_1, \ldots, X_n$ be iid random variables with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for every $\epsilon > 0$,

- Strong law of large number (SLLN):

$$\mathrm{P}(\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon) = 1$$

- Weak law of large number (WLLN):

$$\lim_{n\to\infty} \mathrm{P}(|\bar{X}_n - \mu| < \epsilon) = 1$$

- Central limit theorem (CLT):

as $n \to \infty$,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

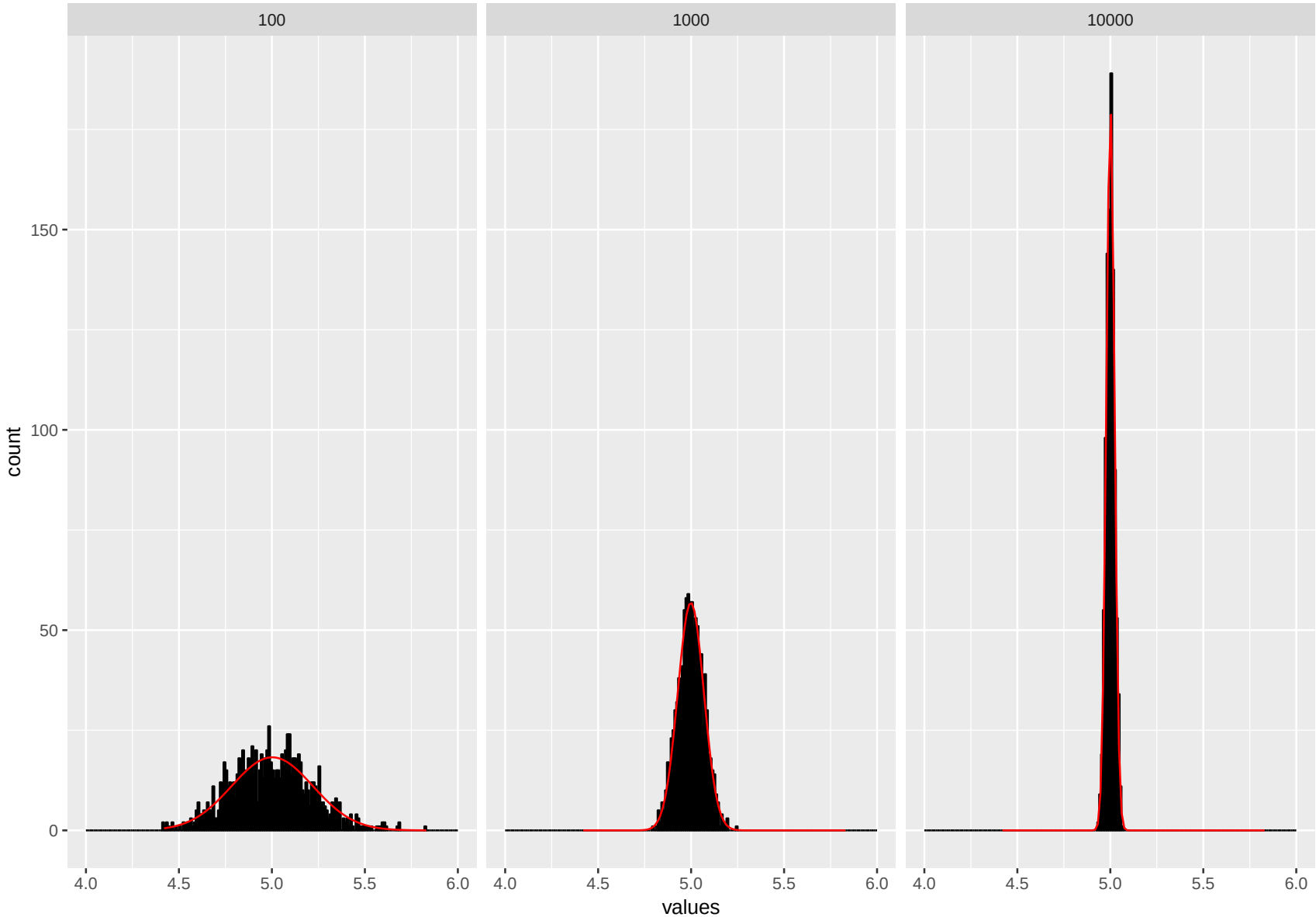
What you need to know

A rigourious treatment of the theory to derive the sampling distribution of $\bar{X}$ is beyond the scope of this course. But you should be able to use the above to justiy yourself in practice.

- SLLN is a stronger result than WLLN; SLLN implies WLLN, but not vice versa.
- SLLN says that with large enough n, the sample mean of random variables with finite variance (can be relaxed to finite expectation) converges **almost surely** (with probability 1) to a constant (or μ).
- WLLN says that with large enough n, **the probability** of the sample mean of random variables with finite variance being really close to a constant (or μ) **converges to 1**.
- CLT states that with large enough n, the distribution of sample mean of random variables with finite variance is approximately normal.

An example

A random sample from Poi($\lambda = 5$), with sample size at $n = 100, 1000$, and 10000.



Central limit theorem (how big should n be?)

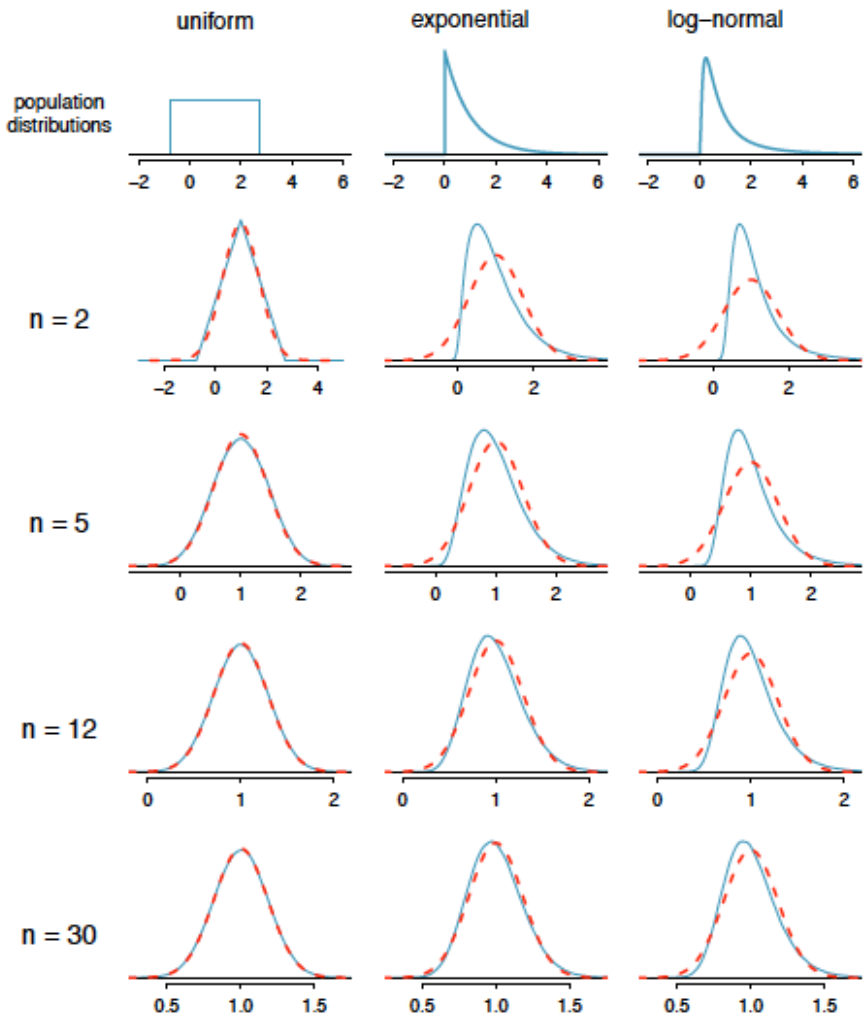


Figure 4.20: Sampling distributions for the mean at different sample sizes and for three different distributions. The dashed red lines show normal distributions.

2

[2] OpenIntro, page 195

Sampling from a normal distribution

Some theoretical results

Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Define two statistics $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$. Show that

1. $\bar{X}$ and S^2 are independent (not tested)
2. $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
3. $(n-1)S^2/\sigma^2$ has a chi-squared distribution with $n-1$ degrees of freedom ($\chi^2(n-1)$) (not tested)

Sketch Proof of a).

The proof can be completed in three steps,

1. show that S^2 can be expressed as a function only of $(X_2 - \bar{X}, \ldots, X_n - \bar{X})$
2. show that $\bar{X}$ is independent of $(X_2 - \bar{X}, \ldots, X_n - \bar{X})$ by writing out the joint p.d.f as a product of the joint p.d.f of $(X_2 - \bar{X}, \ldots, X_n - \bar{X})$ and p.d.f of $\bar{X}$.
3. conclude from the fact that if two random variables are independent, then so are any measurable functions of them.

You only need to be able to show 1), but you should be able to show 2) from STA247 (remember how to do change of variables). 3) is beyond the scope of this course.

Proof of a) Step 1)

Hint:

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

and expressed

$$S^2 = \frac{1}{n-1}\Big[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\Big] \tag{19}$$

$$= \frac{1}{n-1}\Big[\sum_{i=1}^{n}(X_i - \bar{X})^2\Big] \tag{20}$$

as a sum of two things.

Proof of b).

Calculate the moment generating function (STA247) of $\bar{X}$ and compare with the m.g.f of a normal random variable to conclude it is indeed normal with mean μ and variance $\frac{\sigma^2}{n}$.

Proof of b). (Cont'd)

Recall definition of a moment generating function:

$$
\begin{aligned}
M_{\bar{X}}(t) &= \mathrm{E}(e^{t\bar{X}}) && (21)\\
&= \mathrm{E}(e^{t\frac{1}{n}\sum_{i=1}^{n} X_i}) && (22)\\
&= \mathrm{E}(\Pi_{i=1}^{n} e^{t\frac{1}{n}X_i}) && (23)\\
&= \Pi_{i=1}^{n}\mathrm{E}(e^{t\frac{1}{n}X_i}) && (24)\\
&= \Pi_{i=1}^{n}(e^{t\mu/n+1/(2)\sigma^2(t/n)^2}) && (25)\\
&= (e^{t\mu/n+1/(2)\sigma^2(t/n)^2})^n && (26)\\
&= e^{t\mu+1/(2n)\sigma^2 t^2} && (27)
\end{aligned}
$$

Proof of c).

$$(n-1)S^2/\sigma^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2 \tag{28}$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2 - n(\bar{X} - \mu)^2/\sigma^2 \tag{29}$$

- Given that the $\frac{X_i-\mu}{\sigma} \sim \mathcal{N}(0,1)$, we can show by m.g.f or a direct change of variable that $\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(1)$ (this applies to $n(\bar{X} - \mu)^2/\sigma^2$ as well)
- And the fact that the sum of independent chi-squared random variables still follow a chi-squared distribution with d.f. to be the sum of individual d.f's. We have $\sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2 \sim \chi^2(n)$
- Combining the two, we have the result $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$

An example

- From b) of the theorem above, we know that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

- But what about when the variance is unknown?

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

An example (cont'd)

Deriving the sampling distribution of $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$.

- We first construct this statistic to be a ratio of two statistics: $\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)(\sigma/\sqrt{n})}{\sqrt{S^2/n^2}}$
- Notice the top one follows standard normal and the bottom is $\sqrt{\chi^2_{n-1}/(n-1)}$.
- As we have shown in a) that these two are independent, we can thus derive the distribution by looking at the joint distribution of the two components.
- The result you need to know is that $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ follows a student's t-distribution with degrees of freedom $k = n-1$, or $T \sim t_k$

$$f_T(t|k) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \frac{1}{(k\pi)^{1/2}} \frac{1}{(1+t^2/k)^{\frac{k+1}{2}}}, \quad -\infty < t < \infty$$

Finding estimators for a parameter

- Method of moments
- Maximum likelihood (principles of data reduction)
- Bayes estimators (will not be on the test)

What is an estimator and an estimate?

- Recall that a statistic is defined as $T(X_1, \ldots, X_n)$
- An "estimator" or "point estimate" is a statistic used to infer the value of an unknown parameter (so it is always used in reference to a parameter)
- Notation: $\hat{\theta}_n = \hat{\theta}(\mathbf{X}) = T(X_1, \ldots, X_n)$.
- For example, $\bar{X}$ is an estimator of μ.
- An "estimate" is simply the realized value of the estimator, i.e. the estimator (or statistic) evaluated at the actual sample values $(x_1, \ldots, x_n)$.
- Notation: $\hat{\theta}(\mathbf{x}) = t(x_1, \ldots, x_n)$.
- For example, $\bar{x}$ is an estimate of μ.

How do we find estimators?

- In many cases the choice is intuitive, for example, sample mean is a reasonable estimate of the population mean.
- In situation when it's less clear, we need more principled approaches to find estimators.
- Remember, the estimates contain information about the sample while the parameters contain information about the population.
- We need to bridge these two using rigorious statistical methods.

Method of moments

Let $X_1, \ldots, X_n$ be an iid sample from a population with some probability function (p.d.f or p.m.f). Then we can compute the first k sample moments and equate them to each of the first k population moments to find estimators:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i; \quad \mu_1(\theta) = E(X)$$

$$\vdots$$

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k; \quad \mu_k(\theta) = E(X^k)$$

The estimators are found by solving these equations simultaneously.

An example of a normal random sample

Suppose an iid sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, we know the first two sample moments

$$m_1 = \bar{X}$$

and

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

while the population moments

$$E(X) = \mu$$

and

$$E(X^2) = \sigma^2 + \mu^2$$

Thus, the method of moments estimators are $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Method of moments (pro and cons)

- probably the oldest method to find estimator (credit to Karl Pearson in late 1800s)
- simple to implement and always yield some estimate
- estimators might not be plausible
- a good starting point

Likelihood

Let $X_1, \ldots, X_n$ be an iid sample from a population with some probability function (p.d.f or p.m.f). Define the likelihood function (a function of the parameter)

$$L(\theta|\mathbf{x}) = L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \Pi_{i=1}^{n} f(x_i|\theta)$$

- Observe the symmetry between $L(\theta|\mathbf{x})$ and $f(\mathbf{x}|\theta)$.
- The likelihood is a function of the parameter given the data ($\mathbf{x}$)
- The probability is a function of the data given the parameter (θ)
- You can vary the values of θ such that the observed sample is most likely.

Maximum likelihood

Define the maximum likelihood estimator (MLE) $\hat{\theta}$ to be:

$$\hat{\theta}(\mathbf{X}) = \text{argmax}_\theta L(\theta|\mathbf{X})$$

- If $L(\theta|\mathbf{X})$ has a global maximum, then there is a unique $\hat{\theta}$.
- Suppose $L(\theta|\mathbf{X})$ is differentiable with respect to θ, we can find MLE by solving
$$\frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}) = 0; \;\; i = 1, \ldots, k$$
- Then you should verify the solutions are indeed the only extreme point in the interior range of θ by checking the second derivative.
- If θ is restricted to a range, you should also check the boundary points.

An example of a normal random sample

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be iid. The likelihood is then

$$L(\mu, \sigma^2|\mathbf{x}) = \Pi_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{30}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}} \tag{31}$$

$$\frac{\partial}{\partial \mu} L(\mu, \sigma^2|\mathbf{x}) = 2 \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} = 0$$

implies that

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

An example of a normal random sample (cont'd)

On the other hand,

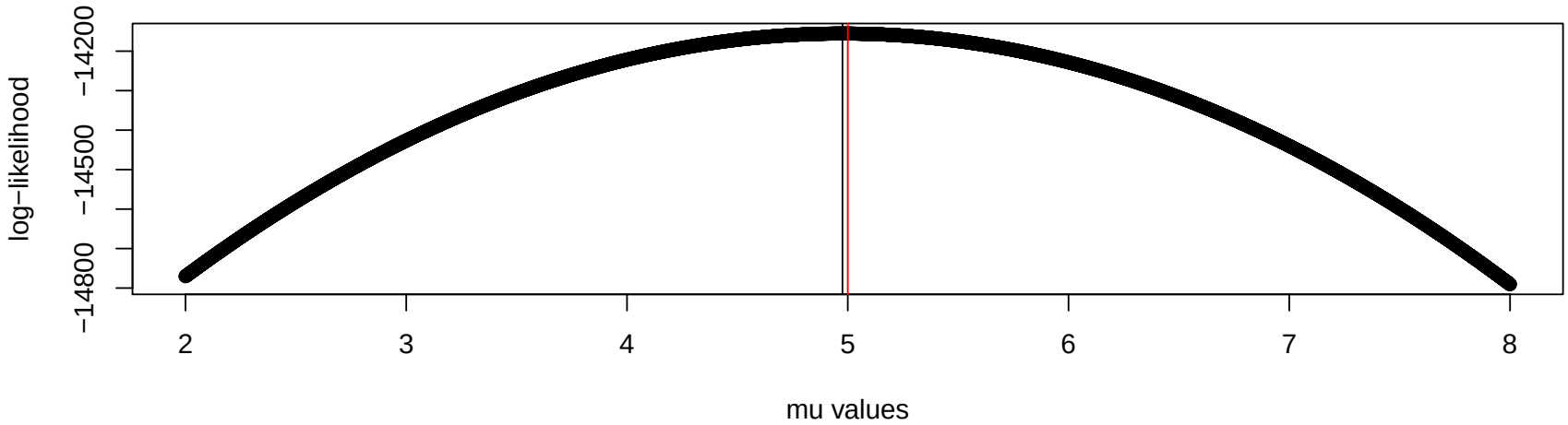
$$\frac{\partial}{\partial \sigma^2} L(\mu, \sigma^2 | \mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^4} = 0$$

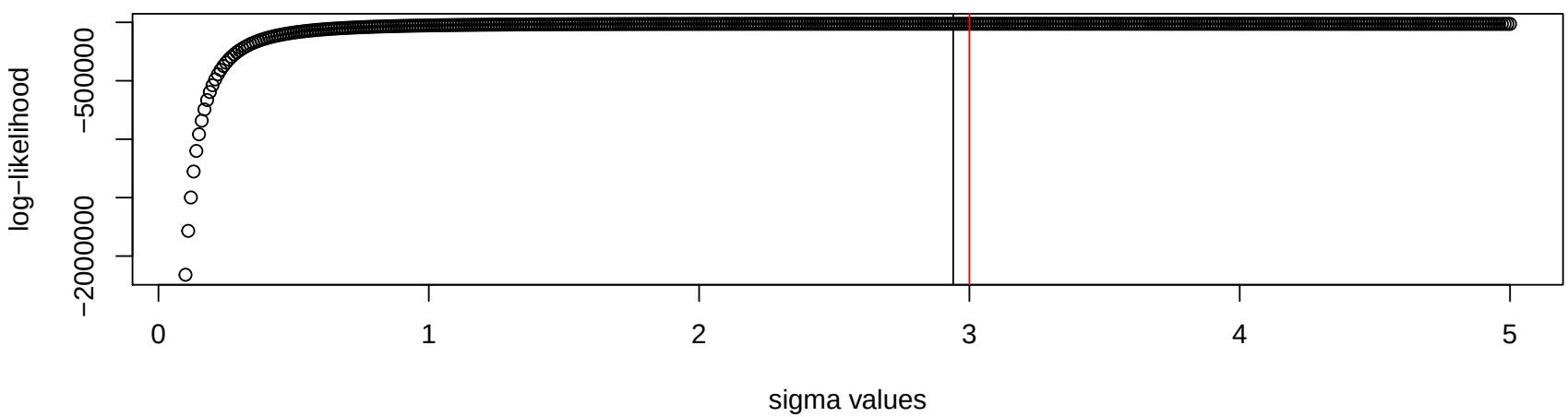
and plugging in $\hat{\mu}$ we have

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n} = \frac{n-1}{n} S^2$$

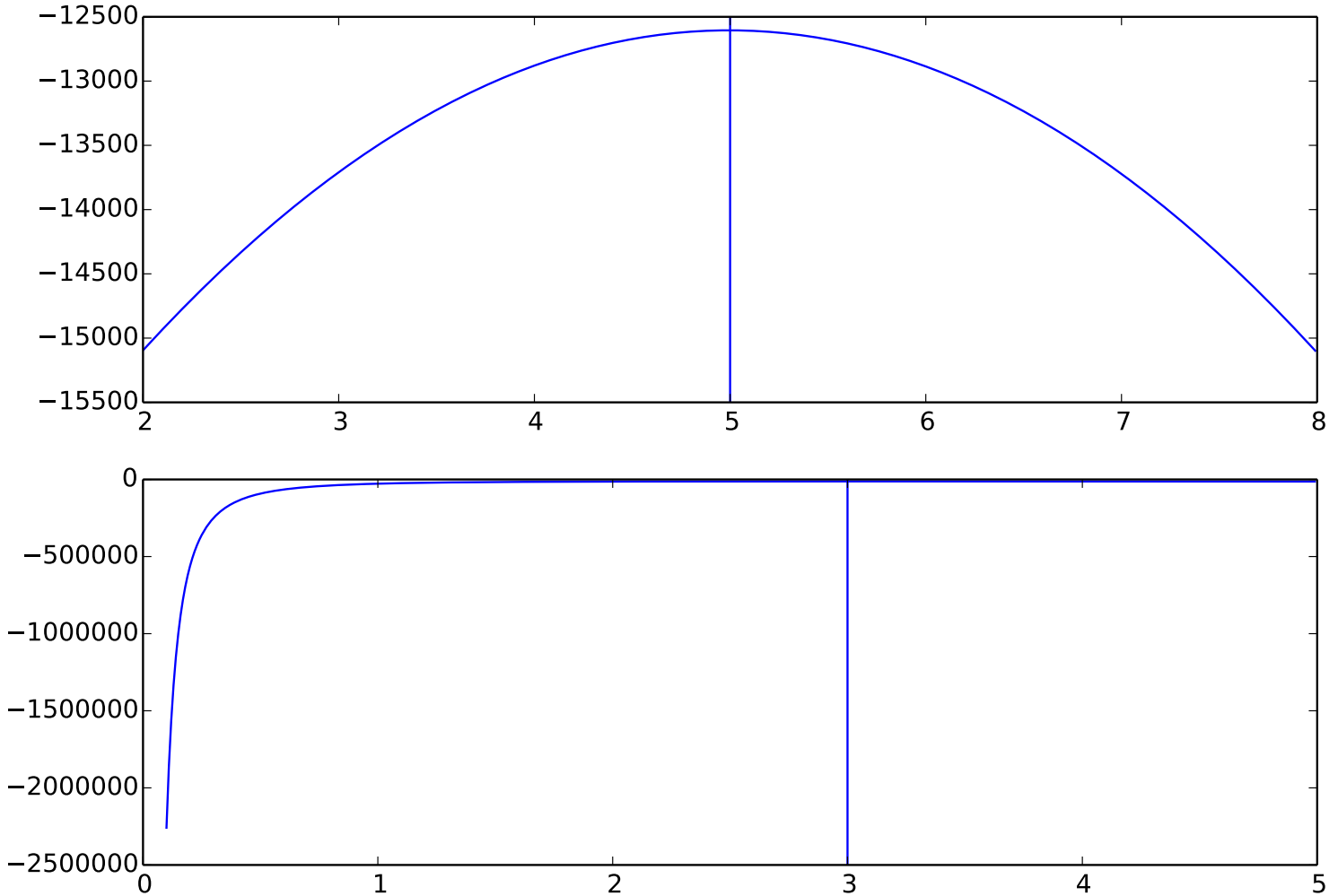
Check for yourself that the two second partial derivatives are negative.

A visual of the MLE (R)





A visual of the MLE (Python)



Evaluating estimators

- Unbiasedness
- Sufficiency
- Consistency

Unbiasedness

An estimator $\hat{\theta}$ is unbiased for the parameter θ if

$$\mathsf{E}_{x|\theta}(\hat{\theta}) = \theta$$

otherwise the difference between the two

$$\mathsf{E}_{x|\theta}(\hat{\theta}) - \theta$$

is called the bias of $\hat{\theta}$ relative to θ.
Examples of unbiased statistics:

- $\bar{X}$ is an unbiased estimator of μ
- S^2 (defined earlier with $1/(n-1)$) is an unbiased estimator of σ^2

(Exercise: show the above).

A common used measure of quality: Mean squared error

MSE of an estimator $\hat{\theta}$ of a parameter θ is a function of θ:

$$\mathrm{E}_{X|\theta}(\hat{\theta} - \theta)^2 = \mathrm{Var}_{X|\theta}(\hat{\theta}) + (\mathrm{E}_{X|\theta}\hat{\theta} - \theta)^2 = \mathrm{Var}_{X|\theta}(\hat{\theta}) + (\mathrm{Bias})^2$$

- both variability of the estimator (precision) and the bias (accuracy) play a role in the mean squared error
- analytically tractable
- A good estimator should have both small - clearly unbiased estimator has some advantage in terms of accuracy.

Exercise

For iid $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, find the MSE of the unbiased estimator $\bar{X}$ and S^2.
Know how to get to these results:

$$\mathrm{E}(\bar{X} - \mu)^2 = \frac{\sigma^2}{n}$$

and

$$\mathrm{E}(S^2 - \sigma^2)^2 = \frac{2\sigma^4}{n-1}$$

Sufficiency

A statistic $T(\mathbf{X})$ is sufficient for θ if the conditional probability distribution of $\mathbf{X}$ does not depend on θ given knowledge of $T(\mathbf{X})$.

- The idea of sufficiency arise from the idea that the statistic contains all information about θ in this sample
- Factorization theorem: Given the probability function is $f_{X|\theta}(x)$, $T(X)$ is sufficient for θ if and only if there exists non-negative functions g and h such that
$$f_{X|\theta}(x) = h(x)g_\theta(T(x))$$
- You should be able to show for the normal case with known variance parameter σ^2, $\bar{X}$ is sufficient for μ.

Consistency

Consider the estimators $\hat{\theta}_n$, consistency refers to the property concerning when $n \to \infty$.
A sequence of estimators $\hat{\theta}_n$ is consistent for parameter θ, if for every $\epsilon > 0$

$$\lim_{n \to \infty} \mathrm{P}(|\hat{\theta}_n - \theta| < \epsilon) = 1.$$

- Consistency means that as sample size approaches infinity, the estimator eventually will be close to the parameter with high probability (not probability 1, so this is weak - as in the WLLN).
- This is different from the previous properties that were restricted to finite samples

Connection between consistency and unbiasedness

Chebyshev's inequality (STA247):

$$\mathrm{P}(|\hat{\theta}_n - \theta| > \epsilon) \leq \frac{\mathrm{E}_{X|\theta}(\hat{\theta}_n - \theta)^2}{\epsilon^2} = \frac{\mathrm{Var}_{X|\theta}(\hat{\theta}_n)}{\epsilon^2} + \frac{\mathrm{Bias}^2}{\epsilon^2}$$

- Thus, if $\hat{\theta}_n$ is unbiased, then as $\lim_{n\to\infty} \mathrm{Var}_{X|\theta}(\hat{\theta}_n) = 0$, $\hat{\theta}_n$ is also consistent.
- You should know that the MLEs are consistent estimators (no need to show).