Stat 231 Chapter 4: Estimation

Statistics

David Duan, 2019 Fall

Contents

Notation. We will use the following notations throughout this chapter.

- $Y_1, \ldots, Y_n$: random variables representing *potential observations* in a random sample.
- $y_1, \ldots, y_n$: actual (observed) data points, realization of random variables $Y_1, \ldots, y_n$.
- μ: mean of the Gaussian distribution, parameter of interest.
- σ: standard deviation of the Gaussian distribution, parameter of interest.

1 Statistical Models and Estimation

In choosing a model for data collected in an empirical study in the analysis of PPDAC, we actually need to consider two probability models:

1. **Attribute Model**: A model for variation in the population or process being studied which includes the attributes which are to be estimated.
2. **Sampling Model**: A model which takes into account how the data were collected and which is constructed in conjunction with the model in (1).

We use these two models to estimate the unknown attributes in the population or process based on the observed data and to determine the uncertainty in these estimates.

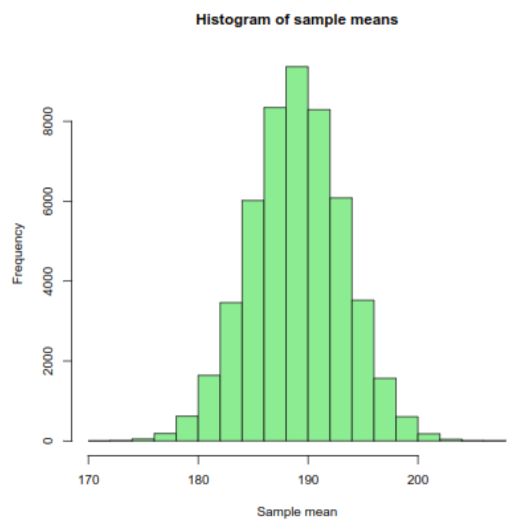
To check the adequacy of a chosen model, we could (1) compare a relative frequency histogram of observed data with the p.d.f. of the assumed model, (2) compare observed frequencies with expected frequencies calculated using the assumed model, (3) compare the empirical c.d.f. with the c.d.f. of assumed model, or (4) examine a Gaussian Q-Q plot.

2 Estimators and Sampling Distributions

We've seen (in Chapter 2) how to choose a model, estimate parameters, and check model fit. We are now going to investigate properties of the estimates, or more precisely, the *process* by which we obtain estimates. For this, we need to think about the idea of *repeated sampling*.

2.1 Repeated Sampling

Let $\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$ be the maximum likelihood estimate of μ, the population mean of a Gaussian distribution. Note that $\hat{\mu} = \bar{y}$ depends on the specific sample we take and would vary as we take samples repeatedly.



Assuming taking samples is a random event, we can think of our estimate for sample mean $\hat{\mu} = \bar{y}$ as a realization of a random variable $\overline{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i$. Our sample mean is therefore a probability distribution that itself has a mean and a variance. We now provide a formal statement.

Prop. 2.1.1 Assume $Y_i \sim G(\mu, \sigma)$ for $i = 1, \ldots, n$ are independent and let $y_1, \ldots, y_n$ be data observed. We could estimate the unknown population mean μ using the maximum likelihood estimate $\hat{\mu} = \bar{y}$. Moreover, there is a random variable associated with sample mean $\overline{y}$:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \iff \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim G\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

We will introduce point estimate and point estimator and generalize the relationship between $\overline{y}$ and $\overline{Y}$ in the next section.

2.2 Point Estimate and Point Estimator

Recall the following definition from Chapter 2.

Def. 2.2.1 A *point estimate* of θ is a function $\hat{\theta} = g(y_1, \ldots, y_n)$ of the observed data used to estimate the unknown parameter θ.

Since estimates vary as we take repeated samples, we can associate with the point estimate $\hat{\theta} = g(y_1, \dots, y_n)$ a random variable $\tilde{\theta} = g(Y_1, \dots, Y_n)$. For example, the random variable associated with $\hat{\theta} = \overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$ is $\tilde{\theta} = \overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.

Def. 2.2.2 A *point estimator* is a random variable which is a function $\tilde{\theta} = g(Y_1, \dots, Y_n)$ of the random variables $Y_1, \dots, Y_n$.

We can view an estimator as a rule that tells us how to process the data to obtain an estimate of an unknown parameter θ, i.e., the numerical value $\hat{\theta} = g(y_1, \dots, y_n)$ is the value obtained using the rule $\tilde{\theta}$ for a particular observed dataset $y_1, \dots, y_n$.

Notation. We use $\hat{\theta}$ to denote the *point estimate* and $\tilde{\theta}$ to denote the *point estimator* for θ.

2.3 Sampling Distribution of an Estimator

Since $\tilde{\theta}$ is a random variable, it has a distribution. In other words, if $\tilde{\theta}$ is a discrete random variable, then it has a probability function; if $\tilde{\theta}$ is a continuous random variable, then it has a probability density function.

Def. 2.3.1 The distribution of an estimator $\tilde{\theta}$ is called its *sampling distribution.*

Knowing the sampling distribution of our estimator, we can then answer questions such as *What is the probability that I will draw a sample that will result in a point estimate $\hat{\mu}$ that is within 1 unit of the true mean μ?*

We restate the second part **Prop 2.1.1** in terms of sampling distribution of an estimator:

Prop. 2.3.2 In general, if $Y_i \sim G(\mu, \sigma)$ for $i = 1, \dots, n$, then the sampling distribution of our estimator $\tilde{\mu} = \overline{Y}$ is given by

$$\overline{Y} \sim G\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Remark. The probability we draw a sample that yields an estimate $\hat{\mu}$ that is close to μ

- ... increases as n increases,
- ... decrease as σ increases,
- ... does not change with μ.

Since the probability does not depend on μ, we can compute it exactly if we know σ and n!

2.4 Gaussian Data with Known Standard Deviation

Prop. 2.4.1 In general, if we have $Y_i \sim G(\mu, \sigma)$, using estimator $\tilde{\mu} = \overline{Y}$, we can transform the problem into a standard normal. That is,

$$
\begin{aligned}
P(|\tilde{\mu}-\mu| \le m) &= P(\mu - m \le \overline{Y} \le \mu + m) \\
&= P\left(\frac{-m}{\sigma/\sqrt{n}} \le \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \le \frac{m}{\sigma/\sqrt{n}}\right) \\
&= P\left(\frac{-m\sqrt{n}}{\sigma} \le Z \le \frac{m\sqrt{n}}{\sigma}\right)
\end{aligned}
$$

where $Z \sim G(0,1)$.

Ex. 2.4.2 For $m = 0.1$, $\sigma = 0.5$, and $n = 49$, $P(|\tilde{\mu}-\mu| \le 0.1) = P(-1.4 \le Z \le 1.4) = 0.838$. That is, with a sample size of 49, we would expect our sample estimate $\hat{\mu}$ to be within 0.1 unit of the true μ 83.8% of the time.

To summarize, when sampling from a Gaussian distribution to estimate population mean μ, we can make use of the result that the sampling distribution of our estimator $\tilde{\mu} = \overline{Y}$ follows a Gaussian distribution

$$\overline{Y} \sim G\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

The probability we draw a sample that results in a point estimate $\hat{\mu}$ that is within a given distance m of the true value μ depends on σ and n but not μ. Thus, if we know or are prepared to specify σ, we can compute these probabilities directly.

2.5 Non-Gaussian Data

Recall the Central Limit Theorem from Stat 230:

Thm. 2.5.1 (CLT) Let $Y_1, \dots, Y_n$ be independent and identically distributed (i.i.d.) random variables with $E[Y_i] = \mu$ and $Var(Y_i) = \sigma^2$ for $i = 1, \dots, n$. Define

$$Z_n = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}.$$

Then for sufficiently large n, Z_n has an approximately $G(0,1)$ distribution.

In other words, if we have random variables that are independent and identically distributed, then given our sample size is large enough, we can take observations from any probability distribution and transform then into a standard normal distribution!

Ex. 2.5.2 For binomial data with n trials and y successes, the estimator $\tilde{\theta} = Y/n$ has $E(\tilde{\theta}) = \theta$ and $Var(\tilde{\theta}) = \theta(1-\theta)/n$. By the Normal approximation to the Binomial, we have

$$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \sim N(0,1) \quad \text{approximately.}$$

We could use this, for example, to determine how large n should be to ensure that $P(-0.03 \le \tilde{\theta} - \theta \le 0.03) \ge 0.95$ for all $\theta \in [0,1]$.

2.6 Standard Deviation of the Sampling Distribution

The following results should be quite intuitive.

Prop. 2.6.1 We have seen that $sd(\overline{Y}) \approx \sigma/\sqrt{n}$.

1. A larger sample size n will decrease $sd(\overline{Y})$ and more of our sample estimates will be close to the true value μ.
2. A small population standard deviation σ will decrease $sd(\overline{Y})$ and more of our sample estimates will be close to the true value μ.
3. The shape of our distribution will affect how many of our sample estimates will be close to the true value μ, but predicting how is trickier.
4. The true mean μ does not affect how many of our sample estimates are close to the true μ.

3 Interval Estimation

Suppose that for a certain population we are interested in estimating θ using point estimate $\hat{\theta} = g(y_1, \dots, y_n)$ given observed data $y_1, \dots, y_n$. To quantify the uncertainty in our estimate, we try to give an interval of values for θ which are "supported" by the data. Intuitively, if we could see the sampling distribution, we could tell which values are more/less plausible for μ.

3.1 Likelihood Intervals

Recall the following definition from Chapter 2.

Def. 3.1.1 The *relative likelihood function* is defined as

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}, \quad \theta \in \Omega$$

where $0 \leq R(\theta) \leq 1$ for all $\theta \in \Omega$ and $R(\hat{\theta}) = 1$.

Given this definition, we can do things like obtaining an interval of value for the unknown parameter which are "reasonable" given the observed data.

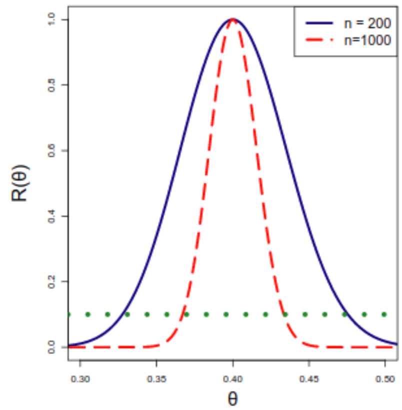
Def 3.1.2 A $100p\%$ *likelihood interval* for the parameter θ is the set $\{\theta : R(\theta) \geq p\}$.

The intuition is that values of θ that result in large values of $L(\theta)$ (and hence $R(\theta)$, because the denominator for $R(\theta)$ is a constant) are more plausible. For example, if $R(\theta_0) = 0.5$, then the data are half as likely if $\theta = \theta_0$ than if $\theta = \hat{\theta}$, the MLE of θ. We can think of a 50% likelihood interval as being the values of θ for which the data are at worst half as likely as they would be if $\theta = \hat{\theta}$.

Remark. The set $\{\theta : R(\theta) \geq p\}$ is not necessarily an interval unless $R(\theta)$ is unimodal, but this is the case for all models that we will consider in this course.

Prop 3.1.3 We now provide some general guidelines for interpreting a likelihood interval.

- Values of θ inside a 10% likelihood interval are plausible given the observed data.
- Values of θ inside a 50% likelihood interval are very plausible given the observed data.
- Values of θ outside a 10% likelihood are implausible given the observed data.
- Values of θ outside a 1% likelihood interval are very implausible given the observed data.



As the sample size n increases, the graph of the relative likelihood function $R(\theta)$ becomes more "concentrated" around θ. Consequently, likelihood intervals becomes narrower as the sample size increases. Both these statements reflect the fact that larger data sets contain more information about the unknown parameter θ.

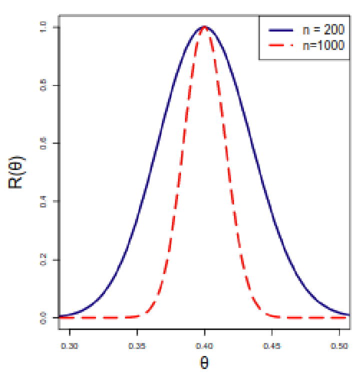
3.2 Log Relative Likelihood Function

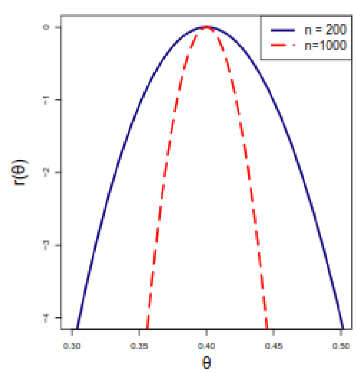
Recall the following definition from Chapter 2.

Def. 3.2.1 The *log relative likelihood function* is given by $r(\theta) = \log[R(\theta)] = \ell(\theta) - \ell(\hat{\theta})$ for $\theta \in \Omega$ where $\ell(\theta) = \log[L(\theta)]$.

Remark.

1. The maximum value of $r(\theta)$ is $\log(1) = 0$. (e.g., the right diagram below.)





2. If $R(\theta)$ is unimodal, then $r(\theta)$ is unimodal, and both graphs attain their maximum value at the MLE of θ.
3. However, $R(\theta)$ and $r(\theta)$ differ in shape: $R(\theta)$ looks bell-shaped while $r(\theta)$ resembles a quadratic function of θ.

Prop. 3.2.2 Since $R(\theta) \geq p \iff r(\theta) \geq \log(p)$, we can plot $r(\theta)$ and draw a line at $r(\theta) = \log(p)$ to obtain an $100p\%$ likelihood interval for θ.

4 Confidence Intervals and Pivotal Quantities

4.1 Interval Estimator

Recall that we can view a sample (point) estimator as a realization of a random variable called a (point) estimator, because point estimates will vary depending on the sample we take. Now, likelihood intervals, which are interval estimates of an unknown parameter θ, also varies depending on the sample. This motivates the concept of an "interval estimator".

Def. 4.1.1 An *interval estimator* $[L(Y), U(Y)]$ is a function (a rule) which can be used to construct an interval of plausible values for the unknown parameter θ.

Just like point estimators, we can view an interval estimator as a rule that tells us how to process the data to obtain an interval estimate of an unknown parameter θ, i.e., the interval estimate $[L(y), U(y)]$ is the value obtained using the rule for a particular observed dataset $y = (y_1, \ldots, y_n)$.

Remark. Both $L(Y)$ and $U(Y)$ are random variables; $L(y), U(y)$ are their realizations.

4.2 Coverage Probability

To determine how "good" one interval estimator is, we look at the interval's coverage probability. Intuitively, an interval is good if it has a high probability containing the true (but known) θ.

Def. 4.2.1 The *coverage probability* for the interval estimator $[L(Y), U(Y)]$ is

$$P(\theta \in [L(Y), U(Y)]) = P[L(Y) \leq \theta \leq U(Y)],$$

i.e., the probability that the random interval $[L(Y), U(Y)]$ contains the true value of θ.

Remark. The interval $[L(Y), U(Y)]$ above is a *random interval* as it takes on specific values depending on the sample we draw.

4.3 Confidence Interval

We now introduce one of the most critical concept of Stat 231. To estimate θ, we want to construct an interval with high coverage probability, e.g., 90%, 95%, or 99%. We call these intervals "confidence intervals".

Def. 4.3.1 A $100p\%$ *confidence interval* for a parameter θ is an interval estimate $[L(y), U(y)]$ s.t.

$$P(\theta \in [L(Y), U(Y)]) = P[L(Y) \leq \theta \leq U(Y)] = p,$$

The value p is called the *confidence coefficient* for the confidence interval.

Warning. Since θ is an unknown CONSTANT associated with the population, it is NOT a random variable and therefore does NOT have a distribution. For an observed set of data y, both $L(y)$ and $U(y)$ are all numerical values. Thus, it is NOT VALID to say that the probability that

θ lies in the interval $[L(y), U(y)]$ is equal to p since θ is a constant. This is the most common mistake when interpreting a confidence interval.

Remark. Approach **Def. 4.3.1** with extreme caution!

- Suppose $p = 0.95$. If we draw repeated independent random samples from the same population and each time we construct the interval $[L(y), U(y)]$ based on the observed data y, then this equation tells us that we should expect 95% of these constructed intervals to contain the true but unknown value of θ (and the rest 5% will not contain θ).
- The correct way to use a confidence interval in a sentence: *we are* 95% *confident that the true value of* θ *is contained in the interval we have constructed.*

Ex. 4.3.2 (Gaussian CI) Suppose we want to estimate μ for a Gaussian distribution. Assume $\sigma = 1$. (We will find out what to do when σ is unknown later. Hint: t distribution.) If $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $G(\mu, 1)$ distribution, then

$$\bar{Y} \sim G\left(\mu, \frac{1}{\sqrt{n}}\right).$$

We know that if we draw an observation from this sample, there is a 95% chance it will lie inside

$$\left[\mu - \frac{1.96}{\sqrt{n}}, \mu + \frac{1.96}{\sqrt{n}}\right].$$

In other words,

$$P\left(\mu - \frac{1.96}{\sqrt{n}} \leq \bar{Y} \leq \mu + \frac{1.96}{\sqrt{n}}\right) = 0.95.$$

Rearranging this equation, we have

$$P\left(\mu \in \left(\bar{Y} - \frac{1.96}{\sqrt{n}}, \bar{Y} + \frac{1.96}{\sqrt{n}}\right)\right) = 0.95.$$

Now, if we draw a sample and observe sample mean $\bar{y}$, then

$$\left(\bar{y} - \frac{1.96}{\sqrt{n}}, \bar{y} + \frac{1.96}{\sqrt{n}}\right)$$

is a 95% confidence interval for the unknown mean μ.

Remark. Observe that the width of the confidence interval decreases as n increases. Intuitively, as our sample size increases, we have more certainty about our estimate, which is reflected in a narrower confidence interval.

Remark. We can plug in actual data to get numeric values. For example, if we define $n = 16$ and observe $\bar{y} = 3.4$, then a 95% confidence interval for μ is

$$\left(3.4 - \frac{1.96}{\sqrt{16}}, 3.4 + \frac{1.96}{\sqrt{16}}\right) = [2.91, 3.89].$$

Again, we CANNOT say $P(\mu \in [2.91, 3.89]) = 0.95$ because μ is constant; we can only say that we are 95% confident that the interval $[2.91, 3.89]$ contains the true but unknown value of μ.

We summarize this example into the following proposition.

Prop. 4.3.3 (Gaussian CI) If $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $G(\mu, 1)$ distribution with known standard deviation but unknown mean, then

$$P\left(\mu \in \left(\overline{Y} - \frac{1.96}{\sqrt{n}}, \overline{Y} + \frac{1.96}{\sqrt{n}}\right)\right) = 0.95$$

and

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

is a 95% confidence interval for μ. We say that we are 95% confident that this interval contains the true but unknown value of μ.

4.4 Pivotal Quantity for Confidence Construction

This following definition implies that probability statements such as $P(Q \leq a)$ and $P(Q \geq b)$ depends on a and b but not on θ or any other unknown information.

Def. 4.4.1 A *pivotal quantity* $Q = Q(Y; \theta)$ is a function of the data Y and the unknown parameter θ such that the distribution of the random variable Q is complete known.

$$Q = \frac{\overbrace{\overline{Y}}^{\text{actual data}} - \overbrace{\mu}^{\text{unknown}}}{\sigma/\sqrt{n}} \sim \underbrace{G(0,1)}_{\text{distribution completely known}}.$$

Remark. We say this is *completely known* because there are no unknown parameters in $G(0, 1)$.

The following example shows how to construct a confidence interval using a pivotal quantity.

Ex. 4.4.2 (CI Construction using PQ) Consider **Ex. 4.3.2** again. Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from $G(\mu, \sigma)$ distribution where $E(Y_i) = \mu$ is unknown but $sd(Y_i) = \sigma$ is unknown. The maximum likelihood estimator for μ is $\tilde{\mu} = \overline{Y}$ with sampling distribution

$$\tilde{\mu} = \overline{Y} \sim G\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Observe Q defined below has a completely known distribution, so Q is a pivotal quantity.

$$Q = Q(Y;\mu) = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0,1)$$

We can use pivotal quantities to construct confidence intervals. Take $n = 16$ and $\sigma = 1$, we have

$$P\left(-1.96 \leq \frac{\overline{Y} - \mu}{1/\sqrt{16}} \leq 1.96\right) = 0.95.$$

Rearranging the inequality inside the brackets,

$$P\left(\mu \in \left(\overline{Y} - \frac{1.96}{\sqrt{16}}, \overline{Y} + \frac{1.96}{\sqrt{16}}\right)\right) = 0.95.$$

Thus, if we take a sample and observe a sample mean of $\bar{y}$, then a 95% CI for μ based on our sample will be

$$\left(\overline{y} - \frac{1.96}{\sqrt{16}}, \overline{y} + \frac{1.96}{\sqrt{16}}\right).$$

We summarize the example into the following proposition.

Prop. 4.4.3 (CI Construction Using PQ) In general, we can use a pivotal quantity to construct a confidence interval as follows:

1. Determine numbers a and b such that $P[a \leq Q(Y;\theta) \leq b] = p$.
2. Re-express the inequality $a \leq Q(Y;\theta) \leq b$ in the form $L(Y) \leq \theta \leq U(Y)$, then

   $$p = P[a \leq Q(Y;\theta) \leq b] = P[L(Y) \leq \theta \leq U(Y)] = P(\theta \in [L(Y), U(Y)])$$

   so the coverage probability of interval $[L(Y), U(Y)]$ is equal to p as desired.
3. For observed data y, the interval $[L(y), U(y)]$ is a $100p\%$ CI for θ.

4.5 Pivotal Quantity for Confidence Interval (Gaussian)

We can simplify **Prop. 4.4.3** when the distribution assumed is Gaussian since it's symmetric.

Ex. 4.5.1 (CI Construction Using PQ; Gaussian) Suppose we want to construct a 95% confidence interval for μ of a Gaussian distribution with known σ.

$$Q = Q(Y;\mu) = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0,1)$$

First, we want to find values a and b such that

$$P\left(a \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq b\right) = 0.95.$$

Since Q is a pivotal quantity, we can find a and b for any probability p (0.95 in this case). There is an infinite number of values of a and b satisfying the equation, e.g., $(-1.8, 2.2)$. Since the normal distribution is symmetric, we choose $a = -1.96$ an $b = 1.96$ which gives the narrowest confidence interval. Our confidence interval (for this example) is therefore symmetric about the point estimate $\hat{\mu}$, so we want to find

$$P\left(-a \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq a\right) = 0.95.$$

By normal tables, $P(Z < -1.96) = P(Z > 1.96) = 0.025$ so $a = 1.96$.

Next, using $a = 1.96$ from step 1, we solve the inequality for μ:

$$-1.96 \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq 1.96 \implies \overline{Y} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + 1.96\frac{\sigma}{\sqrt{n}}.$$

Therefore, a 95% confidence interval for μ based on the observed data $y_1, y_2, \ldots, y_n$ is

$$\left(\bar{y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{y} + 1.96\frac{\sigma}{\sqrt{n}}\right).$$

We summarize the example into the following proposition.

Prop. 4.5.2 (CI Construction Using PQ; Gaussian) We can construct a $100p\%$ confidence interval for μ for Gaussian data where σ is known as follows:

1. User normal tables to find a such that $P(-a \leq Z \leq a) = p$ where $Z \sim (0, 1)$, or equivalently, $P(Z \leq a) = (1 + p)/2$.
2. A $100p\%$ confidence interval for μ is then $\bar{y} \pm a(\sigma/\sqrt{n})$.

Remark. Some useful values from normal tables:

- 90%: $a = 1.645$.
- 95%: $a = 1.960$.
- 99%: $a = 2.576$.

Prop. 4.5.3 For a Gaussian distribution, a $100p\%$ confidence interval for μ is of the form

$$\text{point estimate} \pm (\text{distribution table value}) \times sd(\text{estimator}).$$

Such an interval is often called a "two-sided, equal-tailed" confidence interval.

Remark. We will encounter other examples of two-sided, equal-tailed confidence intervals in this course. Also, *not all confidence intervals are symmetric.*

Remark. We can also use R instead of normal tables. Be familiar with the commands as you will be expected to interpret the results on exams.

- The command `pnorm(a, mu, sigma)` will return $P(Y \leq a)$ where $Y \sim G(\mu, \sigma)$. If we don't specify μ and σ, R assumes $\mu = 0$ and $\sigma = 1$.
  - e.g., `pnorm(3.5, 1, 2) = 0.8943502`: $P(Y < 3.5) = 0.8943502$ given $Y \sim G(1, 2)$.
  - e.g., `pnorm(1.644854) = 0.95`: $P(Y < 1.644854) = 0.95$ given $Y \sim G(0, 1)$.
- The command `qnorm(q, mu, sigma)` returns a value a such that $P(Y \leq a) = q$ where $Y \sim G(\mu, \sigma)$. If we don't specify μ and σ, R assumes $\mu = 0$ and $\sigma = 1$.
  - e.g., `qnorm(0.9, 1, 2) = 3.563103`: $P(Y < 3.563) = 0.9$ given $Y \sim G(1, 2)$.
  - e.g., `qnorm(0.95) = 1.644854`: $P(Y < 1.644954) = 0.95$ given $Y \sim G(0, 1)$.
- To remember what they do, `pnorm` tells us a **p**robability and `qnorm` tells us a **q**uantile (actual value).

4.6 Approximate Pivotal Quantities and Confidence Intervals

Note that for most statistical models, it is not possible to find exact pivotal quantities or confidence intervals for θ.

Def. 4.6.1 An *asymptotic* or *approximate pivotal quantities* is a random variable

$$Q_n = Q_n(Y_1, \ldots, Y_n; \theta)$$

such that as $n \to \infty$, the distribution of Q_n ceases to depend on θ or other unknown information.

Ex. 4.6.2 (Approx. CI for Binomial) Recall that for a binomial experiment, $Y \sim Bin(n, \theta)$ and the point estimator of θ is

$$\tilde{\theta} = \frac{Y}{n}.$$

For large n, the approximate sampling distribution of $\tilde{\theta} = Y/n$ is

$$\frac{\tilde{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \sim G(0, 1) \quad \text{approximately}$$

by the Central Limit Theorem. It can also be shown for large n that

$$Q_n = \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \sim G(0, 1) \quad \text{approximately}$$

Note the $\tilde{\theta}$ in the denominator, in contrast to the previous expression.

Now Q_n is an approximate pivotal quantity which can be used to construct approximate confidence intervals for θ. To obtain a 95% confidence interval,

$$0.95 \approx P\left(-1.96 \leq \frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}} \leq 1.96\right)$$

$$0.95 \approx P\left(\tilde{\theta}-1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}} \leq \theta \leq \tilde{\theta}+1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}\right)$$

So an approximate 95% confidence interval for θ is

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

Note the relationship between a point estimate $\hat{\theta}$ and $\bar{y}$ and a point estimator $\tilde{\theta}$ and $\bar{Y}$.

4.7 Sample Size Calculation

We have seen that confidence intervals for a parameter get narrower as the sample size n increases. When designing a study, researchers need to choose a sample size on the basis of:

- How narrow they would like a confidence interval to be, and
- How much they can afford to spend (time and money).

To do this, we carry out a *sample size calculation*.

Ex. 4.7.1 (Sample Size Calculation) Suppose we plan to select n units at random to estimate θ. The approximate 95% confidence interval for θ is given by

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

which has width

$$2 \times \left(1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right).$$

We might specific that we want a 95% confidence interval of width $\leq 2\ell$, i.e.,

$$1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \quad \text{or} \quad n \geq \left(\frac{1.96}{\ell}\right)^2 \hat{\theta}(1-\hat{\theta}).$$

A criterion that's widely used to choose the sample size n large enough so that the approximate 95% confidence interval is no wider than 2×0.03, i.e., choose n such that

$$\left(\frac{1.96}{0.03}\right)^2 \hat{\theta}(1-\hat{\theta}).$$

Because θ is a proportion, we know $0 < \hat{\theta} < 1$, so RHS of

$$\left(\frac{1.96}{0.03}\right)^2 \hat{\theta}(1 - \hat{\theta})$$

takes its largest value when $\hat{\theta} = 0.5$. We can therefore take a worst-case approach by taking

$$n \geq \left(\frac{1.96}{0.03}\right)^2 0.5 \times 0.5 = 1067.1$$

that if $n = 1068$ then the approximate 95% confidence interval for θ will have width less than 0.03 for all values of $\hat{\theta}$.

Remark. "This poll is accurate to within 3 percentage points 19 times out of 20." This really means that the estimate given is the center of an approximate 95% confidence interval $\hat{\theta} \pm \ell$ for which $\ell = 0.03$.

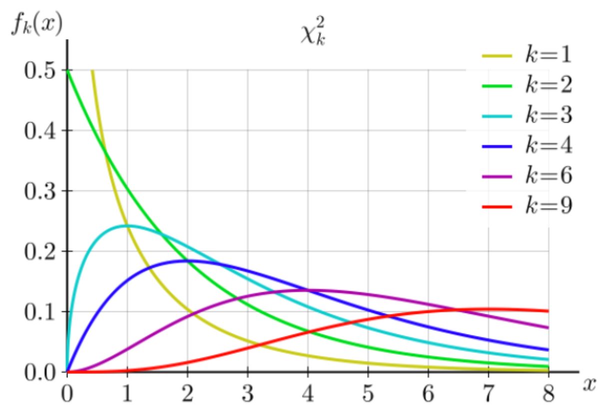
Remark. As an exercise, show that for $\ell = 0.05$ you only need $n = 395$ while for $\ell = 0.02$ you need $n = 2401$. Also think how do these results change if you want a 99% (or 90%) confidence interval?

5 Chi-Squared Distribution

Remark. We only cover key properties for chi-squared distribution in this section. More algebraic details can be found in the course note.

5.1 Properties of Chi-Squared Distribution

Note. 5.1.1 The *chi-squared distribution* is parameterized by its *degrees* of freedom, often denoted k. We would write $Y \sim \chi_k^2$ or $\chi^2(k)$. The value of k affects the shape of the resulting probability density function.



Prop. 5.1.2 If $W_1, W_2, \dots, W_n$ are independent random variables satisfying $W_i \sim \chi_{k_i}^2$, then

$$S = \sum_{i=1}^{n} W_i \sim \chi^2_{\sum k_i}.$$

In other words, the sum of several chi-squared random variables also follows a chi-squared distribution, with degrees of freedom equal to the sum of the degrees of freedom of the component distributions.

Ex. 5.1.3 $W_1 \sim \chi_2^2, W_2 \sim \chi_3^2 \implies W_1 + W_2 \sim \chi_5^2$.

Prop 5.1.4 Chi-squared is related to the standard normal:

- If $Z \sim G(0,1)$, then $Z^2 = W \sim \chi_1^2$. In other words, the square of a standard normal has a chi-squared distribution with 1 degree of freedom.
- If $Z_1, Z_2, \dots, Z_n \sim G(0,1)$, then $S = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$. In other words, the sum of n squared standard normal distributions is chi-squared with n degrees of freedom.

Ex. 5.1.5 If $W \sim \chi_1^2$, then

- $P(W \geq w) = P(Z \geq \sqrt{w}) + P(Z \leq -\sqrt{w}) = 2[1 - P(Z \leq \sqrt{w})]$ where $Z \sim G(0,1)$.
- $P(W \leq w) = P(Z \leq \sqrt{w}) + P(Z \geq -\sqrt{w}) = 2[P(Z \leq \sqrt{w})] - 1$ where $Z \sim G(0,1)$.

Make sure you understand how this is derived! Draw diagrams if necessary.

Prop. 5.1.6 If $W \sim \chi^2_2$, then $W \sim Exp(2)$. In other words, a chi-squared distribution with 2 degrees of freedom is the same as an exponential distribution with parameter 2.

Ex. 5.1.7 For example, $W \sim \chi^2_2 \implies P(W \geq w) = e^{-w/2}$.

This one involves some dirty algebra. If you write out the c.d.f. of χ^2_2, you will see that it is the same as the c.d.f. for exponential(2). Details omitted.

We now show how to get numeric values in a chi-squared distribution. By **Prop. 7.5.1**, a $100p\%$ confidence interval for σ^2 is

$$\left(\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a} \right).$$

where a and b are chosen such that $P(W \leq a) = (1-p)/2$ and $P(W \leq b) = (1+p)/2$.

CHI-SQUARED DISTRIBUTION QUANTILES

99% (0.005 – 0.995), 95% (0.025 – 0.975), 90% (0.05 – 0.95)

df\p	0.005	0.01	0.025	0.05	0.1	0.2	0.8	0.9	0.95	0.975	0.99	0.995
1	0.000	0.000	0.001	0.004	0.016	0.064	1.642	2.706	3.842	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.446	3.219	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.146	1.610	2.343	7.289	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	23.209	25.188
11	2.603	3.054	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	7.807	15.812	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	8.634	16.985	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	9.467	18.151	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	10.307	19.311	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	11.152	20.465	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	12.002	21.615	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.391	10.865	12.857	22.760	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	13.716	23.900	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	14.578	25.038	28.412	31.410	34.170	37.566	39.997
25	10.520	11.524	13.120	14.611	16.473	18.940	30.675	34.382	37.652	40.646	44.314	46.928
30	13.787	14.953	16.791	18.493	20.599	23.364	36.250	40.256	43.773	46.979	50.892	53.672

Some useful numbers:

- $p = 0.9 \implies a = 0.05, b = 0.95.$
- $p = 0.95 \implies a = 0.025, b = 0.975.$
- $p = 0.99 \implies a = 0.005, b = 0.995.$

Remark. We can also use R instead of probability tables. Be familiar with the commands as you will be expected to interpret the results on exams.

- The command `pchisq(w, df)` will return $P(W \leq w)$ where $W \sim \chi^2_{df}$. We must specify df; there is no default.

    - e.g., `pchisq(0.4844186, 4) = 0.025`: $P(W < 0.4844186) = 0.025$ given $W \sim \chi^2_4$.
- The command `qchisq(q, df)` returns a value w such that $P(W \leq w) = q$ where $W \sim \chi^2_{df}$. We must specify df; there is no default.
    - e.g., `qchisq(0.025, 4) = 0.4844186`: $P(W < 0.4844186) = 0.025$ given $W \sim \chi^2_4$.
- To remember what they do, `pchisq` tells us a **p**robability and `qchisq` tells us a **q**uantile (value).

6 Likelihood Intervals and Confidence Intervals

Recall a likelihood interval gives values of θ such that $R(\theta) \geq p$ and a confidence interval gives values of θ such that $P[L(Y) \leq \theta \leq U(Y)] = q$. Both of them give us plausible values of θ but via different methods. We now look at how they relate.

6.1 Likelihood Ratio Statistic

We show that likelihood intervals are also confidence intervals. Recall the *relative likelihood function* is defined as

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}, \quad \theta \in \Omega$$

where $0 \leq R(\theta) \leq 1$ for all $\theta \in \Omega$ and $R(\hat{\theta}) = 1$.

Def. 6.1.1 The random variable

$$\Lambda(\theta) = -2\log\left[\frac{L(\theta)}{L(\tilde{\theta})}\right] = -2\log\left[\frac{L(\theta; Y)}{L(\tilde{\theta}; Y)}\right]$$

where $\tilde{\theta} = \tilde{\theta}(Y)$ is the maximum likelihood estimator is called the *likelihood ratio statistic.*

Prop. 6.1.2 The distribution of $\Lambda(\theta)$ converges to a χ^2_1 distribution as $n \to \infty$.

Thus, we can use $\Lambda(\theta)$ as an approximate pivotal quantity to obtain an approximate confidence interval for θ.

6.2 Likelihood Interval vs. Confidence Interval

Thm. 6.2.1 A $100p\%$ likelihood interval is an approximate $100q\%$ confidence interval where $q = 2P(Z \leq \sqrt{-2\log p}) - 1$ and $Z \sim N(0,1)$.

Proof. A $100p\%$ likelihood interval is defined by $\{\theta : R(\theta) \geq p\}$ which can be rewritten as

$$\left\{\theta : \frac{L(\theta)}{L(\hat{\theta})} \geq p\right\} = \left\{\theta : -2\log\left[\frac{L(\theta)}{L(\hat{\theta})}\right] \leq -2\log p\right\}$$

By **Prop. 6.1.2**, the confidence coefficient for this interval can be approximated by

$$\begin{aligned} P[\Lambda(\theta) \leq -2\log p] &= P\left\{-2\log\left[\frac{L(\theta)}{L(\tilde{\theta})}\right] \leq -2\log p\right\} \\ &\approx P(W \leq -2\log p) && \text{where } W \sim \chi^2_1 \\ &= P(|Z| \leq \sqrt{-2\log p}) && \text{where } Z \sim N(0,1) \\ &= 2P(Z \leq \sqrt{-2\log p}) - 1. && \textbf{Ex. 5.1.5} \quad \square \end{aligned}$$

Ex. 6.2.2 We show that a 10% likelihood interval is an approximate 97% confidence interval.

$$\begin{aligned} q &= 2P(Z \le \sqrt{-2\log(0.1)}) - 1 && \text{where } Z \sim G(0,1) \\ &= 2P(Z \le 2.15) - 1 = 0.96844 \approx 0.97. && \blacksquare \end{aligned}$$

Thm. 6.2.3 If a is a value such that $p = 2P(Z \le a) - 1$ where $Z \sim N(0,1)$, then the likelihood interval $\left\{\theta : R(\theta) \ge e^{-a^2/2}\right\}$ is an approximate $100p\%$ confidence interval.

Proof. The confidence coefficient corresponding to the interval $\left\{\theta : R(\theta) \ge e^{-a^2/2}\right\}$ is

$$\begin{aligned} P\left[\frac{L(\theta)}{L(\tilde{\theta})} \ge e^{-a^2/2}\right] &= P\left\{-2\log\left[\frac{L(\theta)}{L(\tilde{\theta})}\right] \le a^2\right\} \\ &\approx P(W \le a^2) && \text{where } W \sim \chi^2_1 \\ &= 2P(Z \le 1) - 1 && \text{where } Z \sim N(0,1) \\ &= p. && \square \end{aligned}$$

Ex. 6.2.4 We show that a 15% likelihood interval for θ is an $\approx 95\%$ confidence interval for θ.

$$\begin{aligned} 0.95 &= 2P(Z \le 1.96) - 1 && \text{where } Z \sim N(0,1), \\ e^{-(1.96)^2/2} &= e^{-1.9208} \approx 0.1465 \approx 0.15 && \blacksquare \end{aligned}$$

6.3 Approximate Confidence Intervals for Binomial

The intervals are only approximately equivalent and can be numerically quite different. For example, for data y from a binomial distribution with (n, θ), we can obtain an approximate 95% confidence interval using two methods:

- A 15% likelihood interval.
- $\hat{\theta} \pm 1.96\sqrt{\dfrac{\hat{\theta}(1-\hat{\theta})}{n}}$ where $\hat{\theta} = y/n$.

In general, if $\hat{\theta}$ is close to 0.5 or n is large, then the likelihood interval will be fairly accurate about $\hat{\theta}$ and there will be little difference in the two approximate confidence intervals. If $\hat{\theta}$ is close to 0 or 1 and n is not large, however, the likelihood interval will not be symmetric about $\hat{\theta}$ and the two approximate confidence intervals will not be similar. By inspection, you should be able to tell which interval is better supported by the data.

7 Confidence Intervals for Parameters in $G(\mu, \sigma)$

Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $G(\mu, \sigma)$ distribution where μ and σ are both unknown. We use the following estimators for μ and σ:

$$\tilde{\mu} = \overline{Y}$$
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Note that we use sample variance as the estimator for σ^2 instead of $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ because S^2 is unbiased (no overestimate or underestimate), i.e., $E[S^2] = \sigma^2$.

Recall from section 4, if we know σ, we could use the pivotal quantity

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

and derive a 100% confidence interval for μ as

$$\overline{y} \pm a\frac{\sigma}{\sqrt{n}}$$

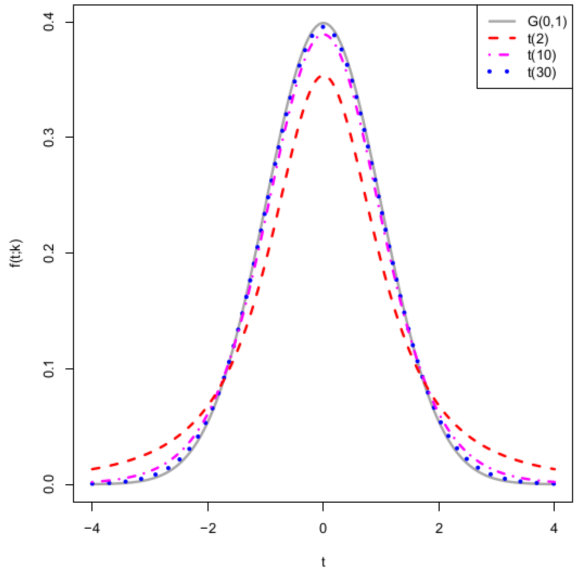
where $P(-a \leq Z \leq a) = p$ and $Z \sim G(0, 1)$.

However, σ is unknown, so we cannot use this result. Luckily, we can simply replace σ with S and obtain the random variable which turns out to also be a pivotal quantity; it has a new distribution: *student's t distribution.*

7.1 Student's t Distribution

Note. 7.1.1 The *student's t distribution* (or simply *t distribution*) has two parameters, t and k. We call k the *degree of freedom* of the distribution.

Notation. We write $T \sim t(k)$ to indicate that the random variable T has a t distribution with k degrees of freedom.



Comparing the p.d.f. for t distribution with that for $G(0,1)$, we see that:

- Both are *unimodal* and *symmetric* about 0.
- For small k, the t distribution has larger tails (i.e., higher values for extreme values).
- For large k, the t distribution is very similar to $G(0,1)$.

As a remark, the t distribution arises as a result of the following proposition.

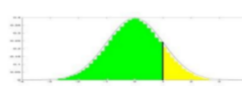
Prop. 7.1.2 Suppose $Z \sim G(0,1)$ and $U \sim \chi_k^2$ independently. Let

$$T = \frac{Z}{\sqrt{U/k}}.$$

Then T has a student's t distribution with k degrees of freedom.

Remark. We can look up t tables given p and df (degree of freedom):

Student t Quantiles



df \ p	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
1	0.3249	0.7265	1.3764	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088	636.6192
2	0.2887	0.6172	1.0607	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271	31.5991
3	0.2767	0.5844	0.9785	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145	12.9240
4	0.2707	0.5686	0.9410	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
5	0.2672	0.5594	0.9195	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
6	0.2648	0.5534	0.9057	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076	5.9588
7	0.2632	0.5491	0.8960	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853	5.4079
8	0.2619	0.5459	0.8889	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008	5.0413
9	0.2610	0.5435	0.8834	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968	4.7809
10	0.2602	0.5415	0.8791	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869
11	0.2596	0.5399	0.8755	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247	4.4370
12	0.2590	0.5386	0.8726	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296	4.3178
13	0.2586	0.5375	0.8702	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520	4.2208
14	0.2582	0.5366	0.8681	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874	4.1405
15	0.2579	0.5357	0.8662	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328	4.0728
					90% CI	95% CI		99% CI		

Remark. We can also use R instead of t tables. Be familiar with the commands as you will be expected to interpret the results on exams.

- The command `pt(t, df)` will return $P(T \leq t)$ where $T \sim t_{df}$. We must specify df; there is no default.
  - e.g., `pt(1.812461, 10) = 0.95`: $P(T < 1.812461) = 0.95$ given $T \sim t_{10}$.
- The command `qt(q, df)` returns a value t such that $P(T \leq t) = q$ where $T \sim t_{df}$. We must specify df; there is no default.
  - e.g., `qt(0.95, 10) = 1.81264`: $P(T < 1.81264) = 0.95$ given $T \sim t_{10}$.
- To remember what they do, `pt` tells us a **p**robability and `qt` tells us a **q**uantile (value).

7.2 Confidence Interval for Gaussian Mean (σ Unknown)

Let's get back to constructing confidence intervals for parameters of a Gaussian distribution.

Prop. 7.2.1 Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $G(\mu, \sigma)$ distribution where neither μ or σ is assumed known. Then

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Observe the LHS contains unknown parameters (namely μ) but its distribution is completely known, so it is a pivotal quantity. We can leverage this to construct confidence intervals for μ without having to assume σ is known!

Thm. 7.2.2 Let $a \in \mathbb{R}$ satisfy $P(-a \leq T \leq a) = p$. The interval

$$\left(\bar{y} - a\frac{s}{\sqrt{n}}, \bar{y} + a\frac{s}{\sqrt{n}}\right)$$

is a $100p\%$ confidence interval for μ.

Proof. The t distribution is symmetric about zero, so we want to look for a value a such that $P(-a \leq T \leq a) = p$, or equivalently,

$$P(T \leq a) = \frac{1+p}{2}$$

where $T \sim t_{n-1}$. Since we are using the pivotal quantity

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

we set T to this and rearrange the expression:

$$P\left(-a \leq \frac{\overline{Y} - \mu}{S/\sqrt{n}} \leq a\right) = p \implies P\left(\overline{Y} - a\frac{S}{\sqrt{n}} \leq \mu \leq \overline{Y} + a\frac{S}{\sqrt{n}}\right) = p.$$

Now, $\bar{y}$ is a realization for $\overline{Y}$, so

$$\left(\bar{y} - a\frac{s}{\sqrt{n}}, \bar{y} + a\frac{s}{\sqrt{n}}\right)$$

as a $100p\%$ confidence interval for μ. □

Remark. It is useful to compare this with the case where σ is known. For a $G(\mu, \sigma)$,

- when σ is known, a $100p\%$ confidence interval is

  $$\bar{y} \pm a_Z \frac{\sigma}{\sqrt{n}}$$

  where $P(Z \leq a_Z) = (1+p)/2$ and $Z \sim G(0,1)$.
- when σ is unknown, a $100p\%$ confidence interval is

  $$\bar{y} \pm a_T \frac{s}{\sqrt{n}}$$

  where $P(T \leq a_T) = (1+p)/2$ and $T \sim t_{n-1}$.

Thus, if $\sigma = s$, then the only difference comes from a_Z vs. a_T.

7.3 Quantifying Uncertainty

We now explore the relationship between individual parameters and the width of a confidence interval. Consider the 95% confidence interval

$$\left(\bar{y} - a\frac{s}{\sqrt{n}}, \bar{y} + a\frac{s}{\sqrt{n}}\right)$$

- *If the confidence level increases to* 99%*, the CI becomes wider,* because wider CI means more likely for the CI to contain the true value.
- *If the sample size increases, then the CI becomes narrower,* because we have obtained more information, so our estimation is more accurate.
- *If the sample standard deviation decreases, then the new CI becomes narrower,* because data points are less likely to be far away from the mean.
- *If the sample mean changes, the new CI has the same width,* because the true mean does not affect the width of a CI.

7.4 Sample Size Calculation Revisited

So far, we've seen two confidence intervals for μ, depending on whether or not we know σ:

$$\bar{y} \pm a\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{y} \pm a\frac{s}{\sqrt{n}},$$

where a is found from either the Z or t_{n-1} distribution. The width, therefore, is

$$2a\frac{\sigma}{n} \quad \text{or} \quad 2a\frac{s}{\sqrt{n}}$$

There is, however, a small problem: if we don't know σ, we need to use the second formula, but then s depends on our sample, so we can't know it ahead of time for a sample size calculation! Moreover, there is no "worst-case" value for s, because the larger it is, the wider our confidence interval will be! Thus, for sample size calculation, we will assume σ is known and use the formula

$$\overline{y} \pm 1.96\frac{\sigma}{\sqrt{n}} \implies n \approx \left(\frac{1.96\sigma}{\ell}\right)^2.$$

In practice, since we usually don't know σ, we would choose n larger than $(1.96\sigma/\ell)^2$.

7.5 Confidence Interval for Gaussian Variance & Standard Deviation

A confidence interval for σ can help inform future sample size calculations.

Recall the point estimator for σ^2:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

It can be shown that

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Observe the random variable Q is a function of data Y_i and the unknown parameter σ whose distribution is completely known. Thus, we got another a pivotal quantity which can help us construct confidence intervals for σ and σ^2.

Prop. 7.5.1 To construct a $100p\%$ confidence interval for σ^2 when μ is unknown,

1. Determine a and b such that $P(a \le W \le b)$ where $W \sim \chi^2_{n-1}$. Since the chi-squared distribution is not symmetric, we find a and b such that
   a. $P(W \le a) = (1-p)/2$, and
   b. $P(W > b) = (1-p)/2$, or equivalently, $P(W \le b) = (1+p)/2$.
2. Re-express the inequality into an interval form. Since $P(a \le W \le b) = p$ with $W \sim \chi^2_{n-1}$ and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

   we have

$$P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) = p \implies P\left(\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right) = p.$$

3. It follows that a $100p\%$ confidence interval for σ is

$$\left(\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right).$$

Remark. (Important) This confidence interval is not symmetric about s^2, the point estimator of σ^2. This means it is not the narrowest possible interval for a given confidence interval!

Cor. 7.5.2 A $100p\%$ confidence interval for σ is

$$\left(\sqrt{\frac{(n-1)s^2}{b}}, \sqrt{\frac{(n-1)s^2}{a}}\right).$$