Statistics

MATH2089

Semester 1, 2018 – Lecture 8

This lecture

7. Inferences concerning a mean

Additional reading:
Sections 5.6 (pp. 234-235), 7.2, 7.3 (pp. 303-306), 7.4 in the textbook (2nd edition)
Sections 5.6 (pp. 238-239), 7.2, 7.3 (pp. 307-311), 7.4 in the textbook (3rd edition)

Confidence interval on the mean of a distribution, variance unknown

Previously we showed how to build confidence intervals for the mean μ of a distribution, assuming that the population variance σ^2 was known

$\rightarrow$ this is probably not very realistic!

Suppose now that the population variance σ^2 is not known

$\rightarrow$ we can no longer make practical use of the core result

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

However, from the random sample $X_1, X_2, \ldots, X_n$ we have a natural estimator of the unknown σ^2: the **sample variance**

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2,$$

which will provide an estimated sample variance $s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$ upon observation of a sample $x_1, x_2, \ldots, x_n$.

Confidence interval on the mean of a normal distribution, variance unknown

A natural procedure is thus to replace σ with the sample standard deviation S, and to work with the random variable

$$T = \sqrt{n}\,\frac{\bar{X} - \mu}{S}$$

In the case of a normal population, Z was just a standardised version of a normal r.v. $\bar{X}$ and was therefore $\mathcal{N}(0,1)$-distributed

However, T is now a ratio of two random variables ($\bar{X} - \mu$ and S)

$\rightarrow$ T is not $\mathcal{N}(0,1)$-distributed !

Indeed, T cannot have exactly the same distribution as Z, as the approximation of the constant σ by a random variable S introduces some extra variability.

$\rightarrow$ the random variable T varies more in value from sample to sample than Z (i.e. $\mathbb{V}\mathrm{ar}(T) > \mathbb{V}\mathrm{ar}(Z)$)

The Student's *t*-distribution

The first person who realised that replacing σ with an estimation did affect the distribution of Z was William Gosset (1876-1937), a British chemist and mathematician who, in the early 20th century, worked at the Guinness Brewery in Dublin.

Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness brewery, so that Guinness prohibited its employees from publishing any scientific papers regardless of the contained information

$\rightarrow$ Gosset negotiated permission to publish, but without a Guinness affiliation, and using the pseudonym *Student*.

He showed that, in a normal population, the exact distribution of T is the so-called t-distribution with $n-1$ degrees of freedom:

$$T \sim t_{n-1}$$

This distribution is now referred to as **Student's *t*-distribution** (which might otherwise have been Gosset's t-distribution).

The Student's t-distribution

A random variable, say T, is said to follow the Student's t-distribution with ν degrees of freedom, i.e.

$$\boxed{T \sim t_\nu}$$

Its probability density function is given by

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad \to S_T = \mathbb{R}$$

for some integer ν.

Note: the Gamma function is given by

$$\Gamma(y) = \int_0^{+\infty} x^{y-1} e^{-x}\, dx, \qquad \text{for } y > 0$$

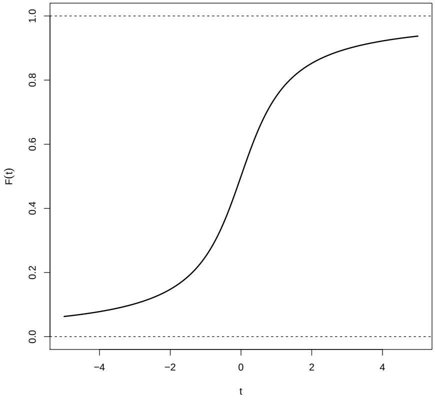
It can be shown that $\Gamma(y) = (y-1) \times \Gamma(y-1)$, so that, if y is a positive integer n,

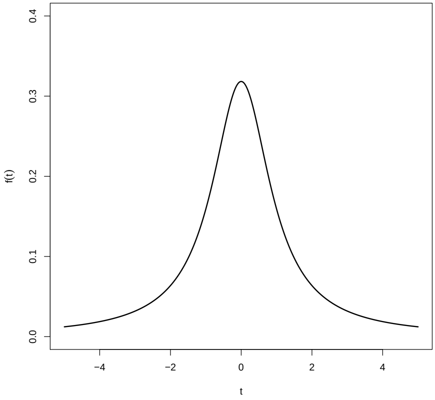
$$\Gamma(n) = (n-1)!$$

There is no simple expression for the Student's t-cdf.

The Student's t-distribution

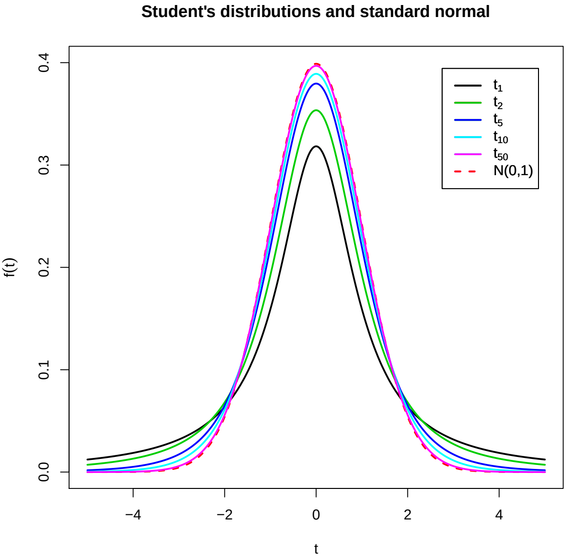
Student's t distribution with 1 degree of freedom





cdf $F(t)$ pdf $f(t) = F'(t)$

The Student's *t*-distribution



The Student's t-distribution

It can be shown that the mean and the variance of the t_ν-distribution are

$$\mathbb{E}(T) = 0 \qquad \text{and} \qquad \mathbb{V}\text{ar}(T) = \frac{\nu}{\nu - 2} \quad (\text{for } \nu > 2)$$

The Student's t distribution is similar in shape to the standard normal distribution in that both densities are symmetric, unimodal and bell-shaped, and the maximum value is reached at 0.

However, the Student's t distribution has heavier tails than the normal

$\rightarrow$ there is more probability to find the random variable T 'far away' from 0 than there is for Z

This is more marked for small values of ν.

As the number ν of degrees of freedom increases, t_ν-distributions look more and more like the standard normal distribution.

In fact, it can be shown that the Student's t distribution with ν degrees of freedom approaches the standard normal distribution as $\nu \to \infty$.

The Student's t-distribution: quantiles

Similarly to what we did for the Normal distribution, we can define the quantiles of any Student's t-distribution:

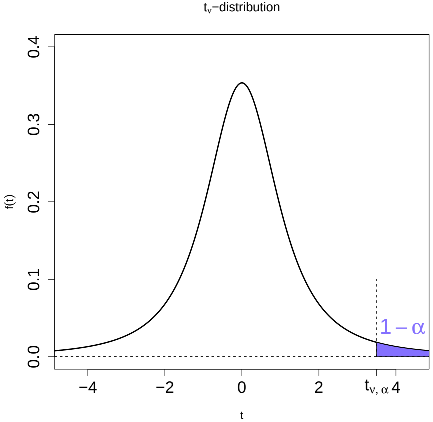
Let $t_{\nu;\alpha}$ be the value such that

$$\mathbb{P}(T > t_{\nu;\alpha}) = 1 - \alpha$$

for $T \sim t_\nu$

Like the standard normal distribution, the symmetry of any t_ν-distribution implies that

$$\boxed{t_{\nu;1-\alpha} = -t_{\nu;\alpha}}$$



$t_{\nu;\alpha}$ is also referred to as t critical value.

Confidence interval on the mean of a normal distribution, variance unknown

So we have, for any $n \geq 2$,

$$T = \sqrt{n}\,\frac{\bar{X} - \mu}{S} \sim t_{n-1}$$

Note: the number of degrees of freedom for the t-distribution is the number of degrees of freedom associated with the estimated variance S^2

It is now easy to find a $100 \times (1-\alpha)\%$ confidence interval for μ by proceeding essentially as we did when σ^2 was known

We may write

$$\mathbb{P}\left(-t_{n-1;1-\alpha/2} \leq \sqrt{n}\,\frac{\bar{X} - \mu}{S} \leq t_{n-1;1-\alpha/2}\right) = 1 - \alpha$$

or

$$\mathbb{P}\left(\bar{X} - t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

t-confidence interval on the mean of a normal distribution

$\rightarrow$ if $\bar{x}$ and s are the sample mean and sample standard deviation of an observed random sample of size n from a normal distribution, a confidence interval of level $100 \times (1-\alpha)\%$ for μ is given by

$$\boxed{\left[\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right]}$$

This confidence interval is sometimes called t-confidence interval, as opposed to $\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ (z-confidence interval)

Because t_{n-1} has heavier tails than $\mathcal{N}(0,1)$, $t_{n-1;1-\alpha/2} > z_{1-\alpha/2}, \forall n$

$\rightarrow$ this reflects the extra variability introduced by the estimation of σ (less accuracy)

Note: One can also define one-sided $100 \times (1-\alpha)\%$ t-confidence intervals

$$\left(-\infty, \bar{x} + t_{n-1;1-\alpha}\frac{s}{\sqrt{n}}\right] \text{ and } \left[\bar{x} - t_{n-1;1-\alpha}\frac{s}{\sqrt{n}}, +\infty\right)$$

t-confidence interval: example

Example

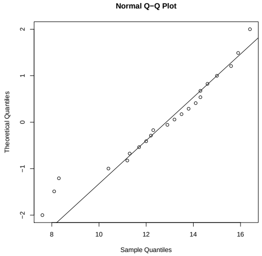
An article in *Materials Engineering* describes the results of tensile adhesion test on 22 $U - 700$ alloy specimens. The load at specimen failure is as follows (in megapascals):

```
7.6, 8.1, 11.7, 14.3, 14.3, 14.1, 8.3, 12.3, 15.9, 16.4,
11.3, 12.0, 12.9, 15.0, 13.2, 14.6, 13.5, 10.4, 13.8,
15.6, 12.2, 11.2
```

Construct a 99% confidence interval for the true average load at failure for this type of alloy. (**Hint:** You can use the Matlab output: $\texttt{tinv}(0.995, 21) = 2.831$)

t-confidence interval: example

The quantile plot below provides good support for the assumption that the population is normally distributed



Confidence interval on the mean of an arbitrary distribution, variance unknown

What if the population is not normal ?

As in the case 'σ^2 known', we can rely on the Central Limit Theorem which asserts that, for n 'large', $Z = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \overset{a}{\sim} \mathcal{N}(0,1)$ to deduce a result like

$$T = \sqrt{n}\,\frac{\bar{X}-\mu}{S} \overset{a}{\sim} t_{n-1}$$

from which we could find a CI on μ for n large enough.

However, recall that, when ν is large, t_ν is very much like $\mathcal{N}(0,1)$

$\rightarrow$ in large samples, estimating σ with S has very little effect on the distribution of T, to which the approximation by the standard normal distribution is more than enough:

$$\boxed{T \overset{a}{\sim} \mathcal{N}(0,1)}$$

Confidence interval on the mean of an arbitrary distribution

Consequently, if $\bar{x}$ and s are the sample mean and standard deviation of an observed random sample of large size n from any distribution, an approximate confidence interval of level $100 \times (1-\alpha)\%$ for μ is

$$\boxed{\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]}$$

This expression holds regardless of the population distribution, as long as n is large enough $\rightarrow$ it is called a large-sample confidence interval.

Generally, n should be at least 40 to use this result reliably (the CLT usually holds for $n \geq 30$, but a larger sample size is recommended because replacing σ by S still results in some additional variability).

As usual, corresponding one-sided confidence intervals could be defined: $(-\infty, \bar{x} + z_{1-\alpha}\frac{s}{\sqrt{n}}]$ and $[\bar{x} - z_{1-\alpha}\frac{s}{\sqrt{n}}, +\infty)$

Confidence interval on the mean: example

Example

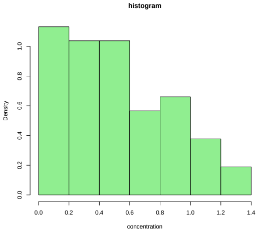
An article in *Transactions of the American Fisheries Society* reports the results of a study to investigate the mercury contamination in largemouth bass. A sample of 53 fishes was selected from some Florida lakes, and mercury concentration in the muscle tissue was measured (in ppm):

`1.23, 0.49, 1.08, ..., 0.16, 0.27`

Find a confidence interval on μ, the mean mercury concentration in the muscle tissue of fish. (**Hint:** You can use the Matlab output: $\texttt{norminv}(0.975) = 1.96$, $\texttt{tinv}(0.975, 52) = 2.007$)

An histogram and a quantile plot for the data are displayed below

Confidence interval on the mean: example



Confidence intervals on the mean: example

For large sample sizes, what if the population is not normal and you still use the t-confidence interval?

Confidence interval on the mean: example

Example

The article “Extravisual Damage Detection? Defining the Standard Normal Tree" (*Photogrammetric Engr. and Remote Sensing, 1981: 515-522*) discusses the use of color infrared photography in identification of normal trees in Douglas fir stands. Among data reported were summary statistics for green-filter analytic optical densitometric measurements on samples of both healthy and diseased trees. For a sample of 69 healthy trees, the sample mean dye-layer density was 1.028, and the sample standard deviation was 0.163. Assume the dye-layer density follows a normal distribution. a) Calculate a 95% two-sided confidence interval for the true average dye-layer density for all such trees. (**Hint:** You can use the Matlab output: $\texttt{norminv}(0.975) = 1.96$, $\texttt{tinv}(0.975, 68) = 1.9955$)

Confidence interval on the mean: example

Example (ctd.)

a) Calculate a 95% two-sided confidence interval for the true average dye-layer density for all such trees. (**Hint:** You can use the Matlab output: $\texttt{norminv}(0.975) = 1.96$, $\texttt{tinv}(0.975, 68) = 1.9955$)

Confidence interval on the mean: example

Example (ctd.)

b) Suppose the investigators had made a rough guess of 0.16 for the value of s before collecting data. What sample size would be necessary to obtain an interval width of 0.05 for a confidence level of 95%?

Confidence intervals for the mean: summary

The several situations leading to different confidence intervals for the mean can be summarised as follows:

The first question is: **Is the population normal?** (check from a histogram and a quantile plot, for instance)

- if yes, is σ known ?
  - if yes, use an exact z-confidence interval:
    $$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$
  - if no, use an exact t-confidence interval:
    $$\left[\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$
- if no, use an approximate large sample confidence interval:
  $$\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right],$$
  (provided the sample size is large, say $n \geq 40$)

What if the sample size is small and the population is not normal ?
$\rightarrow$ check on a case by case basis (beyond the scope of this course)

Prediction interval for a future observation

In some situations, we may be interested in predicting a future observation of a variable.

$\rightarrow$ different than estimating the mean of the variable !

$\rightarrow$instead of confidence intervals, we are after
$100 \times (1 - \alpha)\%$ prediction interval on a future observation

As an illustration, suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and standard deviation σ

$\rightarrow$ we wish to predict the value X_{n+1}, a single future observation

As X_{n+1} comes from the same population as $X_1, X_2, \ldots, X_n$, information contained in the sample should be used to predict X_{n+1}

$\rightarrow$ the **predictor** of X_{n+1}, say X^*, should be a statistic

Prediction interval for a future observation

We desire the predictor to have expected prediction error equal to 0:

$$\mathbb{E}(X_{n+1} - X^*) = 0 \qquad \Longleftrightarrow \qquad \mathbb{E}(X^*) = \mu$$

$\rightarrow$ the predictor X^* must be an unbiased estimator for μ!

We said that an efficient unbiased estimator for μ was the sample mean, so we take it as predictor:

$$X^* = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Now, the variance of the prediction error is

$$\begin{aligned}\mathbb{V}\mathrm{ar}(X_{n+1} - X^*) = \mathbb{V}\mathrm{ar}(X_{n+1} - \bar{X}) &= \mathbb{V}\mathrm{ar}(X_{n+1}) + \mathbb{V}\mathrm{ar}(\bar{X}) \\ &= \sigma^2 + \frac{\sigma^2}{n} = \sigma^2\left(1 + \frac{1}{n}\right)\end{aligned}$$

(because X_{n+1} is independent of $X_1, X_2, \ldots, X_n$ and so of $\bar{X}$)

Prediction interval for a future observation

Finally, because both X_{n+1} and $\bar{X}$ are normally distributed (normal population), the prediction error $X_{n+1} - \bar{X}$ is also normally distributed

Hence,

$$Z = \frac{X_{n+1} - \bar{X}}{\sigma\sqrt{1 + \frac{1}{n}}} \sim \mathcal{N}(0, 1)$$

Replacing the possibly unknown σ with the sample standard deviation S yields

$$T = \frac{X_{n+1} - \bar{X}}{S\sqrt{1 + \frac{1}{n}}} \sim t_{n-1}$$

Manipulating Z and T as we did previously for CI leads to the $100 \times (1 - \alpha)\%$ z- and t-prediction intervals on the future observation:

$$\left[\bar{x} - z_{1-\alpha/2}\,\sigma\,\sqrt{1 + \frac{1}{n}}, \bar{x} + z_{1-\alpha/2}\,\sigma\,\sqrt{1 + \frac{1}{n}}\right]$$

$$\left[\bar{x} - t_{n-1;1-\alpha/2}\,s\,\sqrt{1 + \frac{1}{n}}, \bar{x} + t_{n-1;1-\alpha/2}\,s\,\sqrt{1 + \frac{1}{n}}\right]$$

Prediction interval for a future observation: remarks

Remark 1:

The length of a confidence interval on μ of level $100 \times (1 - \alpha)\%$ is $2 \times z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$.

The length of a prediction interval on X_{n+1} of level $100 \times (1 - \alpha)\%$ is $2 \times z_{1-\alpha/2} \times \sigma\sqrt{1 + \frac{1}{n}}$.

Prediction intervals for a single observation will always be longer than confidence intervals for μ, because there is more variability associated with one observation than with an average.

Remark 2:

As n gets larger ($n \to \infty$), the length of the CI for μ decreases to 0 (we are more and more accurate when estimating μ), but this is not the case for a prediction interval: the inherent variability of X_{n+1} never vanishes, even when we have observed many other observations before!

Prediction interval for a future observation: example

Example

Reconsider the example on Slide 13. Find a 99% confidence interval for the true average load at failure. We plan to test a 23rd specimen. Find a 99% prediction interval on the load at failure for this specimen. (**Hint:** You can use the Matlab output: $\texttt{tinv}(0.995, 21) = 2.831$)

From the data ($n = 22$) we had found $\bar{x} = 12.67$ MPa and $s = 2.47$ MPa, and a 99% confidence interval for μ was $[11.18, 14.16]$

Now, $t_{21;0.995} = 2.831$ (hint), so that a 99% prediction interval for the next observation is

$$\left[\bar{x} \pm t_{n-1;1-\alpha/2}\, s\sqrt{1+\frac{1}{n}}\right] = \left[12.67 \pm 2.831 \times 2.47 \times \sqrt{1+\frac{1}{22}}\right]$$

$$= [5.52, 19.82]$$

$\rightarrow$ we are 99% confident that the failure load for the next specimen will be between 5.52 and 19.82 MPa

Inferences concerning proportions

Many engineering problems deal with proportions, percentages or probabilities:

we are concerned with the proportion of defectives in a lot, with the percentage of certain components which will perform satisfactorily during a stated period of time, or with the probability that a newly produced item meets some quality standards

$\rightarrow$ qualitative information can also be included in statistical studies!

It should be clear that problems concerning proportions, percentages or probabilities are really equivalent: a percentage is merely a proportion multiplied by 100, and a probability is a proportion in a (infinitely) long series of trials.

We would like to learn about π, the **proportion of the population that has a characteristic of interest**, but as usual all we have is just a sample of size *n* from that population

$\rightarrow$ inference about π $\rightarrow$ confidence interval for π

Estimation of a proportion

In this situation, the random variable to study is

$$X = \begin{cases} 1 & \text{if the individual has the characteristic of interest} \\ 0 & \text{if not} \end{cases}$$

which is Bernoulli distributed, with parameter being the value π of interest:

$$X \sim \text{Bern}(\pi)$$

The random sample $X_1, X_2, \ldots, X_n$ is a set of n independent Bern(π) random variables.

$\rightarrow$ the number Y of individuals of the sample with the characteristic is

$$Y = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, \pi)$$

and the **sample proportion** is

$$\hat{P} = \frac{Y}{n}$$

Estimation of a proportion

This sample proportion $\hat{P}$ is obviously a natural candidate for estimating the population proportion π.

From the properties of the Binomial distribution, we know that

$$\mathbb{E}(Y) = n\pi \qquad \text{and} \qquad \mathbb{V}\text{ar}(Y) = n\pi(1-\pi)$$

so that $\mathbb{E}(\hat{P}) = \frac{1}{n}\mathbb{E}(Y) = \pi$ and $\mathbb{V}\text{ar}(\hat{P}) = \frac{1}{n^2}\mathbb{V}\text{ar}(Y) = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$

Hence, $\hat{P}$ is an **unbiased** and **consistent estimator** for π:

$$\mathbb{E}(\hat{P}) = \pi \qquad \text{and} \qquad \mathbb{V}\text{ar}(\hat{P}) = \frac{\pi(1-\pi)}{n} \quad (\to 0 \text{ as } n \to \infty)$$

$\to$ the standard error of $\hat{P}$ is thus $\text{sd}(\hat{P}) = \sqrt{\frac{\pi(1-\pi)}{n}}$

Upon observation of a random sample $x_1, x_2, \ldots, x_n$, in which $y = \sum_{i=1}^{n} x_i$ individuals have the characteristics, an estimate of π is

$$\hat{p} = \frac{y}{n}$$

Sampling distribution

We could make inference about π from $\hat{p}$ using the Binomial distribution of Y. However, it is probably easier to use the **Central Limit Theorem**. Indeed:

$$\hat{P} = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

so that $\hat{P}$ is actually a (particular) sample mean, for which the **CLT** guarantees that

$$\boxed{\sqrt{n}\,\frac{\hat{P}-\pi}{\sqrt{\pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0,1)}$$

if n is 'large' ($\overset{a}{\sim}$ again stands for "approximately follows")

We also know that the quality of the approximation depends on the symmetry of the initial distribution of the X_i's, here Bern(π)

$\rightarrow$ π should not be too close to 0 or 1 $\rightarrow$ empirical rule: $n\hat{p}(1-\hat{p}) > 5$

Confidence interval for a proportion

As the sampling distribution

$$\sqrt{n}\,\frac{\hat{P}-\pi}{\sqrt{\pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0,1)$$

is just a particular case of $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \overset{a}{\sim} \mathcal{N}(0,1)$, we can use (almost) directly the large-sample confidence interval we derived for a mean

Specifically, we have that

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \sqrt{n}\,\frac{\hat{P}-\pi}{\sqrt{\pi(1-\pi)}} \leq z_{1-\alpha/2}\right) \simeq 1-\alpha$$

or

$$\mathbb{P}\left(\hat{P} - z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \leq \pi \leq \hat{P} + z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right) \simeq 1-\alpha$$

$\rightarrow$ a confidence interval for π takes shape

Confidence interval for a proportion

Unfortunately, the standard error of $\hat{P}$, that is the factor $\sqrt{\frac{\pi(1-\pi)}{n}}$, contains the unknown π.

In such a situation, we may replace the unknown value by its estimate, that is, to use the estimated standard error of the estimator

$$\widehat{\text{sd}(\hat{P})} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

in the expression of the confidence interval.

Consequently, if $\hat{p}$ is the sample proportion in an observed random sample of size n, an approximate two-sided confidence interval of level $100 \times (1-\alpha)\%$ for π is given by

$$\boxed{\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]}$$

As this is based on the CLT and requires n 'large', it is a large sample confidence interval for π.

One-sided confidence intervals for a proportion

We may also find one-sided large-sample confidence intervals for the proportion π by a simple modification of the previous development

We find:

$$\boxed{\left[0, \hat{p} + z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,\right]}$$

and

$$\boxed{\left[\,\hat{p} - z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, 1\right]}$$

Choice of the sample size

Since $\hat{p}$ is the estimate of π, we can define the error in estimating π by $\hat{p}$ as $e = |\hat{p} - \pi|$. From Slide 33, we are approximately $100 \times (1-\alpha)\%$ confident that this error is less than

$$z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$$

In situations where the sample size can be selected, we may choose n to be $100 \times (1-\alpha)\%$ confident that the error is less than any specified value e:

$$n = \left(\frac{z_{1-\alpha/2}}{e}\right)^2 \pi(1-\pi) \qquad \text{(compare Slide 26, Lecture 7)}$$

$\rightarrow$ this depends on π, for which no information is available at this point

Idea: use an upper bound which holds for any value of π

Actually, $\pi(1-\pi) \leq 1/4$, with equality for $\pi = 1/2$, thus with

$$\boxed{n = \left(\frac{z_{1-\alpha/2}}{2e}\right)^2}$$

we are at least $100 \times (1-\alpha)\%$ confident that this error is less than e and this, regardless of the value of π (this is very conservative, though).

Confidence interval for a proportion: example

Example

In a random sample of 85 car engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow. a) Find a 95% confidence interval on the true proportion π of produced bearings that exceeds the roughness specification.

Confidence interval for a proportion: example

Example (ctd.)

In a random sample of 85 car engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow. b) How large is a sample required if we want to be 95% confident that the error in estimating π is less than 0.05?

Confidence interval for a proportion: example

Example

The article “Repeatability and Reproducibility for Pass/Fail Data" (*J. of Testing and Eval., 1997: 151-153*) reported that in $n = 48$ trails in a particular laboratory, 16 resulted in ignition of a particular type of substrate by a lighted cigarette. Let π denote the long-run proportion of all such trials that would result in ignition. Find a 95% confidence interval on the true proportion π.

Objectives

Now you should be able to:

- Construct z- and t-confidence intervals on the mean of a normal distribution, advisedly using either the normal distribution or the Student's t distribution □
- Construct large sample confidence intervals on a mean of an arbitrary distribution with unknown variance □
- Explain the difference between a confidence interval and a prediction interval □
- Construct prediction intervals for a future observation in a normal population □
- Construct confidence intervals on a population proportion □

Recommended exercises:

$\rightarrow$ Q7, Q9, p.301, Q13, Q15 p.302, Q20 p.303, Q35 p.319, Q39 p.320, Q43(a-b) p.320, Q55 p.328, (optional) Q71, Q73 p.340, Q55 p.238 (2nd edition)

$\rightarrow$ Q7, Q9, p.305, Q16 p.307, Q21 p.307, Q37 p.324, Q42 p.325, Q46(a-b) p.326, Q58 p.334, (optional) Q75, Q77 p.347, Q57 p.242 (3rd edition)