6 Introduction to formal statistical inference

Formal statistical inference uses probability theory to quantify the reliability of data-based conclusions. We want information on a population. We can use:

for example: true mean fill weight of food jars
true average number of cycles to failure of a kind of spring
true mean breaking strength of a wire rope.

1. Point estimates:

   e.g. sample mean
   For example measure breaking strength of 6 wire ropes as 5, 3, 7, 3, 10, 1

   estimate $\mu \approx \bar{x} = \dfrac{5+3+7+3+10+1}{6} = 4.83$ tons

2. Interval estimates:

   μ is likely to be inside the interval $(4.83-2, 4.83+2) = (2.83, 6.83)$

   We are confident that the true mean breaking strength μ is somewhere in $(2.83, 6.83)$. But how confident?

6.1 Large-sample confidence intervals for a mean

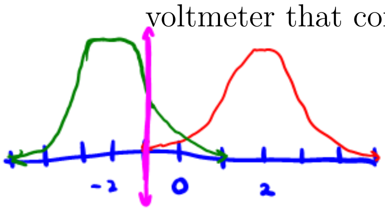
Many important engineering applications of statistics fit the following mold. Values for parameters of a data-generating process are unknown. Based on data, the goal is

1. identify an interval of values likely to contain an unknown parameter
2. quantify "how likely" the interval is to cover the correct value.

Definition 6.1. A *confidence interval* for a parameter (or function of one or more parameters) is a data-based (1) interval of numbers thought likely to contain the parameter (or function of one or more parameters) possessing a (2) stated probability-based confidence or reliability.

A confidence interval is a realization of a **random interval**, an interval on the real line with a random variable at one or both of the endpoints.

Example 6.1 (Instrumental drift)**.** Let Z be a measure of instrumental drift of a random voltmeter that comes out of a certain factory. Say $Z \sim N(0,1)$. Define a random interval:



$$(Z-2, Z+2)$$

endpoints are random variables

What is the probability that -1 is inside the interval?

$$P(-1 \text{ is in } (Z-2, Z+2)) = P(Z-2 < -1 < Z+2)$$
$$= P(Z-1 < 0 < Z+3)$$
$$= P(-1 < -Z < 3)$$
$$= P(-3 < Z < 1)$$
$$= \Phi(1) - \Phi(-3)$$
$$= 0.84$$

Example 6.2 (More practice). Calculate:

1. $P(2 \text{ in } (X-1, X+1))$, $X \sim N(2, 4^{=2^2})$

$$P(2 \in (X-1, X+1)) = P(X-1 < 2 < X+1)$$
$$= P(-1 < 2-X < 1)$$
$$= P(-1 < X-2 < 1)$$
$$= P(-\tfrac{1}{2} < Z < \tfrac{1}{2}) \qquad Z \sim N(0,1)$$
$$= \Phi(\tfrac{1}{2}) - \Phi(-\tfrac{1}{2})$$
$$= 0.6915 - 0.3085$$
$$= 0.383$$

2. $P(6.6 \text{ in } (X-2, X+1))$, $X \sim N(7, 2)$ $\leftarrow (\sqrt{2})^2$

$$P(6.6 \in (X-2, X+1)) = P(X-2 < 6.6 < X+1)$$
$$= P(-2 < 6.6 - X < 1)$$
$$= P(-1 < X - 6.6 < 2)$$
$$= P(-1.4 < X-7 < 1.6)$$
$$= P\left(-\frac{1.4}{\sqrt{2}} < Z < \frac{1.6}{\sqrt{2}}\right) \qquad Z \sim N(0,1).$$
$$\approx P(-.99 < Z < 1.13)$$
$$= \Phi(1.13) - \Phi(-.99)$$
$$= 0.8708 - .1611 = .7097$$

Example 6.3 (Abstract random intervals). Let's say $X_1, X_2, \ldots, X_n$ are iid with $n \geq 25$, mean μ, variance σ^2. [We can find a random interval that provides a lower bound for μ with $1-\alpha$ probability] for $\alpha \in (0,1)$

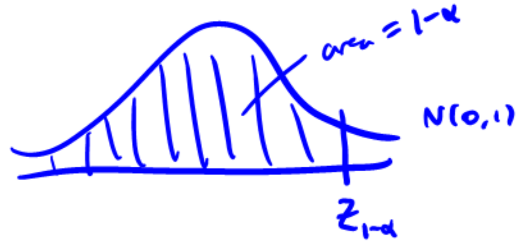
Want A s.t. $P(\mu \in (A, \infty)) = 1-\alpha$

We know $\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$ by CLT

$$\Rightarrow \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1) \quad \text{(standardization)}.$$

Let $z_{1-\alpha}$ denote the $1-\alpha$ quantile of $N(0,1)$ distribution.

$$\Rightarrow P(Z \leq z_{1-\alpha}) = 1-\alpha$$



$$\Rightarrow P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha}\right) \approx 1-\alpha$$

$$P\left(\bar{X} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \leq \mu\right) \approx 1-\alpha$$

$$\text{i.e. } P\left(\mu \in \left(\underbrace{\bar{X} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}}_{\text{"}A\text{"}}, \infty\right)\right) \approx 1-\alpha.$$

Calculate:

1. $P(\mu \in (-\infty, \overline{X} + \underset{\substack{\uparrow \\ 1-\alpha \text{ quantile of } N(0,1)}}{z_{1-\alpha}}\frac{\sigma}{\sqrt{n}})), X_i \sim N(\mu, \sigma^2)$

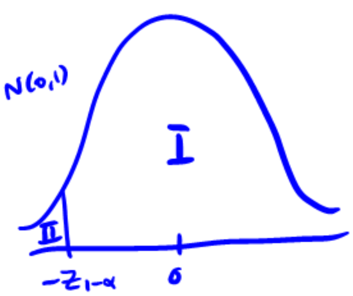
$= P(\mu < \bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}})$

$= P(-z_{1-\alpha}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu)$

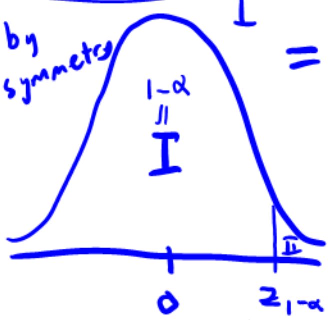
$= P(-z_{1-\alpha} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}})$

$\approx \underbrace{P(-z_{1-\alpha} < Z)}_{\mathrm{I}}$ $\quad Z \sim N(0,1)$ by CLT (if $n \geq 25$)

$= 1-\alpha$ (by symmetry).



by symmetry



$z_{1-\alpha} = 1-\alpha$ quantile $\Rightarrow P(Z < z_{1-\alpha}) = 1-\alpha$ (def'n)

2. $P(\mu \in (\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})), X_i \sim N(\mu, \sigma^2)$

$= P(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$

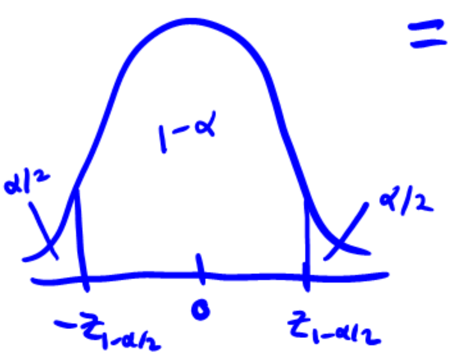
$= P(-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu - \bar{X} < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$

$= P(-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$

$= P(-z_{1-\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2})$.

$\approx P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2})$ for $Z \sim N(0,1)$ by CLT (assuming $n \geq 25$)

$= 1-\alpha$



$1 - \frac{\alpha}{2} = P(Z \leq z_{1-\alpha/2})$

6.1.1 A Large-n confidence interval for μ involving σ

A $1-\alpha$ **confidence interval** for an unknown parameter is the realization of a random interval that contains that parameter with probability $1-\alpha$.

called the "confidence level"

For random variables $X_1, X_2, \ldots, X_n$ iid with $\mathrm{E}(X_1) = \mu$, $\mathrm{Var}(X_1) = \sigma^2$, a $1-\alpha$ confidence interval for μ is

come from data "realization" of $\bar{X}$

$$(\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$$

which is a **realization** from the random interval

$$(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

- Two-sided $1-\alpha$ confidence interval for μ

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

or written as $\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$

- One-sided $1-\alpha$ confidence interval for μ with a upper confidence bound

$$\left(-\infty, \bar{x} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right)$$

- One-sided $1-\alpha$ confidence interval for μ with a lower confidence bound

$$\left(\bar{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right)$$

Example 6.4 (Fill weight of jars)**.** Suppose a manufacturer fills jars of food using a stable filling process with a known standard deviation of $\sigma = 1.6$g. We take a sample of $n = 47$ ≥ 25 jars and measure the sample mean weight $\overline{x} = 138.2$g. A two-sided 90% confidence interval ($\alpha = 0.1$) for the true mean weight μ is:

$(1-\alpha)100\%$

$(1-\alpha)100 = 90$

$1-\alpha = .9 \Rightarrow \alpha = .1$

$$\left(\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

.95 quantile of the standard normal distribution

$$= \left(138.2 - z_{1-.1/2=.95}\frac{1.6}{\sqrt{47}}, 138.2 + z_{.95}\frac{1.6}{\sqrt{47}}\right)$$

$$= (138.2 - 1.64(0.23), 138.2 + 1.64(0.23))$$

$$= (137.82, 138.58)$$

Could have also written as 138.2 ± 0.38 g

Interpretation:

We are $\overset{(1-\alpha)}{90\%}$ confident that the $\overbrace{\text{true mean}}^{\mu}$ $\overbrace{\text{fill weight}}^{\text{context specific}}$ is between $\underbrace{137.82\text{g and } 138.58}_{\text{endpoints}}$ $\underset{\text{units}}{\text{g}}$.

If we took 100 more samples of 47 jars each, roughly 90 of those samples would yield a confidence interval containing the true mean fill weight, μ.

What if we just want to be sure that the true mean fill weight is high enough?

We could use a one-sided 90% CI with a lower bound.

$$\left(\bar{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}},\ \infty\right)$$

$$= \left(138.2 - z_{.9}\frac{1.6}{\sqrt{47}},\ \infty\right)$$

$$= \left(138.2 - 1.28(0.23),\ \infty\right)$$

$$= (137.91,\ \infty)$$

We are 90% confident that the true mean fill weight is above 137.91 g.

Example 6.5 (Hard disk failures)**.** F. Willett, in the article "The Case of the Derailed Disk Drives?" (*Mechanical Engineering*, 1988), discusses a study done to isolate the cause of link code A failure in a model of Winchester hard disk drive. For each disk, the investigator measured the breakaway torque (in. oz.) required to loosen the drive's interrupter flag on the stepper motor shaft. Breakaway torques for $\overset{n=}{26}$ disk drives were recorded, with a sample mean of 11.5 in. oz. Suppose you know the true standard deviation of the breakaway torques is 5.1 in. oz. Calculate and interpret:

σ"

1. A two-sided 90% confidence interval for the true mean breakaway torque of the relevant type of Winchester drive.

$\sigma = 5.1, \quad \bar{x} = 11.5, \quad n = 26, \quad 1-\alpha = .9 \Rightarrow \alpha = .1$

$$\left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$
$$= \left(11.5 - z_{.95}\frac{5.1}{\sqrt{26}},\ 11.5 + z_{.95}\frac{5.1}{\sqrt{26}}\right)$$
$$= \left(11.5 - 1.64(1.0002),\ 11.5 + 1.64(1.0002)\right)$$
$$= (9.86, 13.14)$$

We are 90% confident that the true mean breakaway torque lies between 9.86 in. oz and 13.14 in. oz.

2. An analogous two-sided 95% confidence interval.

$1-\alpha = .95 \Rightarrow \alpha = .05$

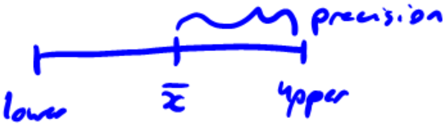
$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 11.5 \pm z_{.975}\frac{5.1}{\sqrt{26}}$$
$$= 11.5 \pm 1.96(1.0002)$$
$$= (9.54, 13.46)$$

We are 95% confident that the true mean breakaway torque lies between 9.54 in. oz. and 13.47 in. oz.

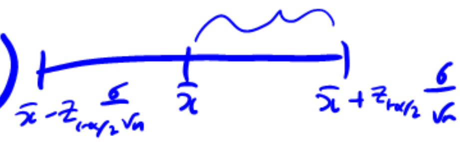
Note: as confidence levels $(1-\alpha)$ increase, the confidence interval gets wider.

Example 6.6 (Width of a CI)**.** If you want to estimate the breakaway torque with a 2-sided, 95% confidence interval with ± 2.0 in. oz. of precision, what sample size would you need?

interval precision = interval half width



two-sided 95% CI: $\left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$



precision $= z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$

$\Rightarrow$ we want $z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq 2$

i.e. $z_{.975}\frac{5.1}{\sqrt{n}} \leq 2$

$$1.96\frac{5.1}{\sqrt{n}} \leq 2$$

$$\frac{9.996}{\sqrt{n}} \leq 2$$

$$n \geq 24.98$$

$$\Rightarrow n \geq 25$$

We would need a sample of at least 25 disks to have at least a precision of 2 in oz.

6.1.2 A generally applicable large-n confidence interval for μ

Although the equations for a $1-\alpha$ confidence interval is mathematically correct, it is severely limited in its usefulness because

it requires us to know σ. It is unusual to have to estimate μ and know σ in real life.

If $n \geq 25$ and σ is *unknown*, $Z = \frac{\overline{X}-\mu}{s/\sqrt{n}}$, where

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

is still **approximately standard normally distributed**. So, you can replace σ in the confidence interval formula with the sample standard deviation, s.

- Two-sided $1-\alpha$ confidence interval for μ

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

- One-sided $1-\alpha$ confidence interval for μ with a upper confidence bound

$$\left(-\infty,\ \bar{x} + z_{1-\alpha}\frac{s}{\sqrt{n}}\right)$$

- One-sided $1-\alpha$ confidence interval for μ with a lower confidence bound

$$\left(\bar{x} - z_{1-\alpha}\frac{s}{\sqrt{n}},\ \infty\right)$$

Example 6.7. Suppose you are a manufacturer of construction equipment. You make 0.0125 inch wire rope and need to determine how much weight it can hold before breaking so that you can label it clearly. Here are breaking strengths, in kg, for 41 sample wires:

```
 [1] 100.37  96.31  72.57  88.02 105.89 107.80  75.84  92.73  67.47  94.87
[11] 122.04 115.12  95.24 119.75 114.83 101.79  80.90  96.10 118.51 109.66
[21]  88.07  56.29  86.50  57.62  74.70  92.53  86.25  82.56  97.96  94.92
[31]  62.00  93.00  98.44 119.37 103.70  72.40  71.29 107.24  64.82  93.51
[41]  86.97
```

The sample mean breaking strength is $\bar{x} =$ 91.85 kg and the sample standard deviation is $s =$ 17.6 kg. Using the appropriate 95% confidence interval, try to determine whether the breaking strengths meet the requirement of at least 85 kg. => one-sided CI w/ lowerbound

$1-\alpha = .95 \Rightarrow \alpha = .05$

$\bar{x} = 91.85$

$s = 17.6$

$n = 41$

$$\left(\bar{x} - z_{1-\alpha}\frac{s}{\sqrt{n}}, \infty\right)$$

$$= \left(91.85 - z_{.95}\frac{17.6}{\sqrt{41}}, \infty\right)$$

$$= \left(91.85 - 1.64\left(\frac{17.6}{\sqrt{41}}\right), \infty\right)$$

$$= (87.3422, \infty)$$

With 95% confidence, we have shown that the true mean breaking strength is above 87.3422 kg. Hence, we meet the 85 kg requirement with 95% confidence.

6.2 Small-sample confidence intervals for a mean (Ch 6.3 in the text)

The most important practical limitation on the use of the methods of the previous sections is

the requirement that n must be large ($n \geq 25$)

That restriction comes from the fact that without it,

There is no way (in general) to conclude that $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ is approximately $N(0,1)$ (because we cannot use the CLT).

So, if one mechanically uses the large-n interval formula $\overline{x} \pm z\frac{s}{\sqrt{n}}$ with a small sample,

There is no way of assessing what actual level of confidence should be declared.

If it is sensible to model the observations as iid normal random variables, then we can arrive at inference methods for small-n sample means.

If this is true $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ isn't standard Normal, BUT it is a different named distribution!

6.2.1 The Student t distribution

Definition 6.2. The *(Student) t distribution with degrees of freedom parameter* ν is a continuous probability distribution with probability density

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2} \qquad \text{for all } t.$$

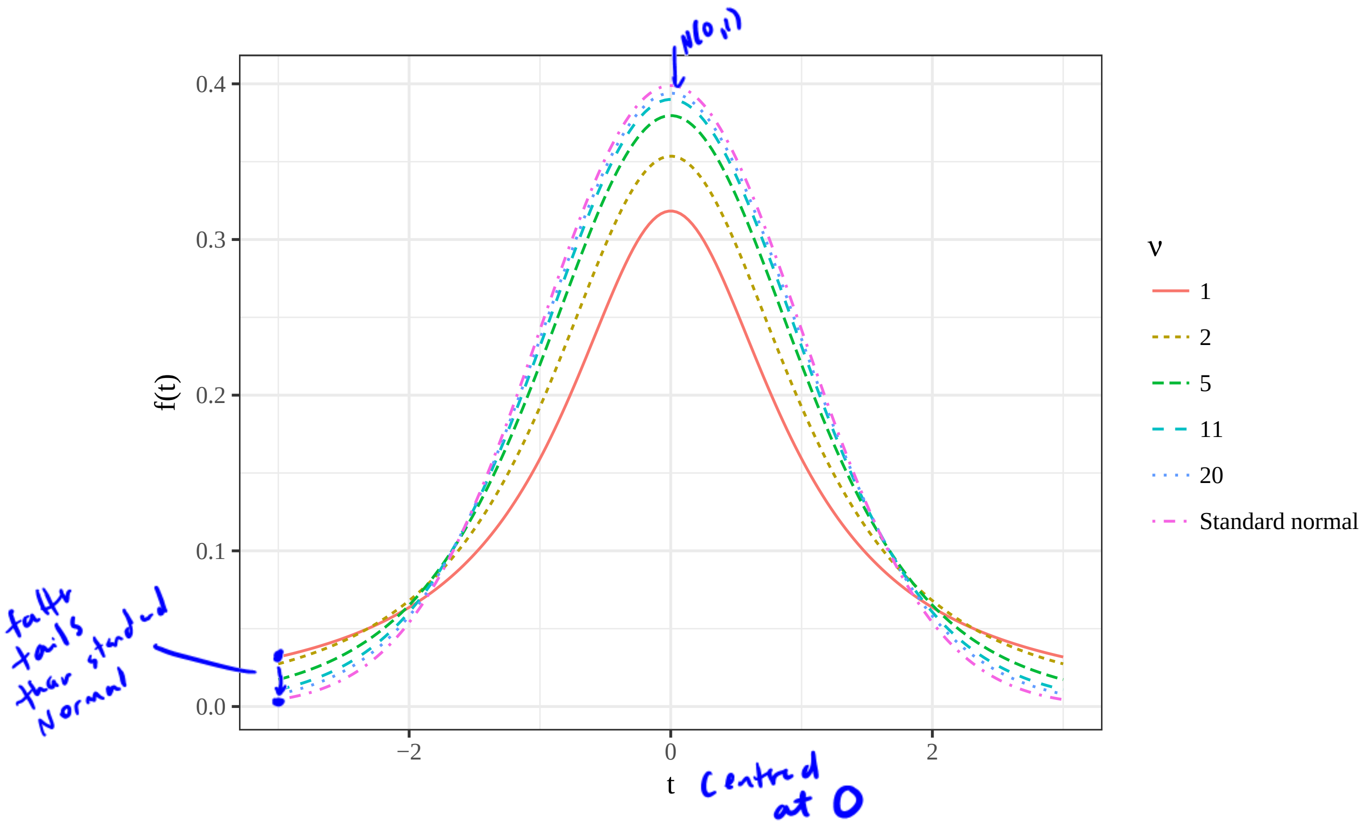
$t \in \mathbb{R}$

The t distribution

- is bell-shaped and symmetric about 0
- has fatter tails than the normal, but approaches the shape of the normal as $\nu \to \infty$.

We use the t table (Table B.4 in Vardeman and Jobe) to calculate quantiles.

(see attached)



definition of the $Q(.9)$

Example 6.8 (t quantiles)**.** Say $T \sim t_5$. Find c such that $P(T \leq c) = 0.9$.

Table B.4
t Distribution Quantiles

ν	$Q(.9)$	$Q(.95)$	$Q(.975)$	$Q(.99)$	$Q(.995)$	$Q(.999)$	$Q(.9995)$
1	3.078	6.314	12.706	31.821	63.657	318.317	636.607
2	1.886	2.920	4.303	6.965	9.925	22.327	31.598
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869

Figure 1: Student's t distribution quantiles.

$P(T \leq 1.476) = 0.9$

$Q(p)$ for a t_ν random variable is denoted $t_{\nu,p}$

so $t_{5,0.9} = 1.476$.

6.2.2 Small-sample confidence intervals, σ unknown

If we can assume that $X_1, \ldots, X_n$ are iid with mean μ and variance σ^2, and are also normally distributed, (even if $n < 25$),

if $n < 25$, we can't use the CLT.

But we know: $\dfrac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$ (since $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$)

We can then use $t_{n-1,1-\alpha/2}$ instead of $z_{1-\alpha/2}$ in the confidence intervals.

Note: df for the t distribution is equal to $\underline{\underline{n-1}}$

- Two-sided $1-\alpha$ confidence interval for μ

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

- One-sided $1-\alpha$ confidence interval for μ with a upper confidence bound

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

- One-sided $1-\alpha$ confidence interval for μ with a lower confidence bound

$$\left(\bar{x} - t_{n-1,1-\alpha}\frac{s}{\sqrt{n}},\ \infty\right)$$

Example 6.9 (Concrete beams). 10 concrete beams were each measured for flexural strength (MPa). Assuming the flexural strengths are iid normal, calculate and interpret a two-sided 99% CI for the flexural strength of the beams.

[1] 8.2 8.7 7.8 9.7 7.4 7.8 7.7 11.6 11.3 11.8

$n=10, \quad \alpha = 0.01$

$\bar{x} = \frac{1}{10}(8.2+8.7+\ldots+11.8) = 9.2$

$s = \sqrt{\frac{1}{9}\left[(8.2-9.2)^2+(8.7-9.2)^2+\ldots+(11.8-9.2)^2\right.} = 1.76$

Two sided 99% CI: $\left(\bar{x} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}, \ \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right)$

$= \left(9.2 - t_{9,0.995}\frac{1.76}{\sqrt{10}}, \ 9.2 + t_{9,0.995}\frac{1.76}{\sqrt{10}}\right)$

$= (9.2 - 3.250(0.556), \ 9.2 + 3.250(0.556))$

$= (7.393, \ 11.007)$

We are 99% confident that the true mean flexural strength of this kind of beam is between 7.393 MPa and 11.007 MPa.

Is the true mean flexural strength below the minimum requirement of 11 MPa? Find out with the appropriate 95% CI. ↳ We need an upper 95% CI.

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

$= \left(-\infty, 9.2 + t_{9,.95}\frac{1.76}{\sqrt{10}}\right)$

$= (-\infty, \ 9.2 + 1.8333(0.556))$

$= (-\infty, \ 10.22)$

We are 95% confident that the true mean flexural strength is below 10.22 MPa. (Notice this is less than 11)

So at $\alpha = .05$, we have shown the true mean flexural strength is below 11 MPa, and the requirement is met.

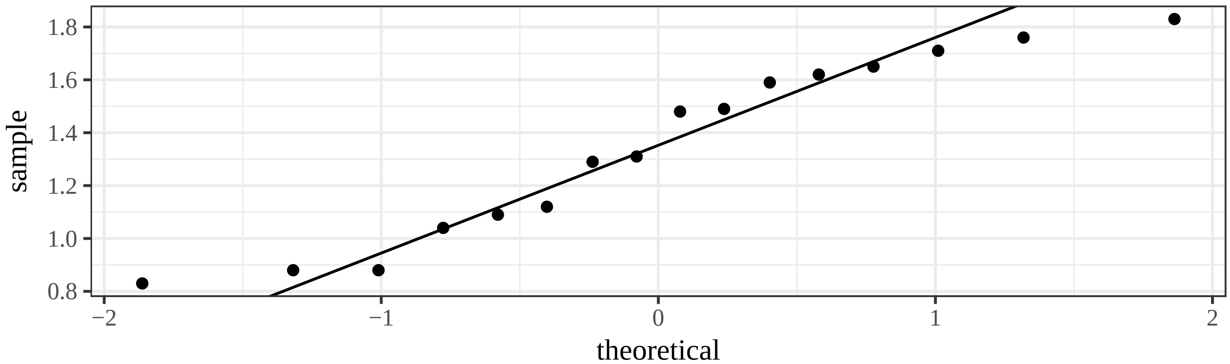
Example 6.10 (Paint thickness)**.** Consider the following sample of observations on coating thickness for low-viscosity paint.

in mm

```
 [1] 0.83 0.88 0.88 1.04 1.09 1.12 1.29 1.31 1.48 1.49 1.59 1.62 1.65 1.71
[15] 1.76 1.83
```

$n = 16$

A normal QQ plot shows that they are close enough to normally distributed.



Calculate and interpret a two-sided 90% confidence interval for the true mean thickness.

$n = 16 \qquad \alpha = 0.1$

$\bar{x} = \frac{1}{16}(0.83 + \ldots + 1.83) = 1.35 \text{ mm}$

$s = \sqrt{\frac{1}{15}\left[(0.83-1.35)^2 + \ldots + (1.83-1.35)^2\right]} = 0.34 \text{ mm}$

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

$= \left(1.35 - t_{15,.95}\frac{0.34}{\sqrt{16}},\ 1.35 + t_{15,.95}\frac{0.34}{\sqrt{16}}\right)$

$= \left(1.35 - 1.75(0.08\overline{5}),\ 1.35 + 1.75(0.085)\right)$

$= (1.201, 1.499)$

We are 90% confident that the true mean thickness is between 1.201 mm and 1.499 mm.

6.3 Hypothesis testing

Last section illustrated how probability can enable confidence interval estimation. We can also use probability as a means to use data to quantitatively assess the plausibility of a trial value of a parameter.

Statistical inference is using data from the sample to draw conclusions about the population.

1. Interval estimation (confidence intervals) estimating population parameters and specifying the degree of precision of the estimate.

2. Hypothesis testing testing the validity of statements about the population that are framed in terms of parameters.

Definition 6.3. Statistical *significance testing* is the use of data in the quantitative assessment of the plausibility of some trial value for a parameter (or function of one or more parameters).

i.e. assess the plausibility of a process mean value of 138g for fill weight of baby food.

Significance (or hypothesis) testing begins with the specification of a trial value (or **hypothesis**).

Definition 6.4. A *null hypothesis* is a statement of the form

$$\underbrace{\text{Parameter}}_{\text{(not a statistic)}} = \#$$

or

$$\text{Function of parameters} = \#$$

fixed number

for some # that forms the basis of investigation in a significance test. A null hypothesis is usually formed to embody a status quo/"pre-data" view of the parameter. It is denoted H_0.

"null" because it is always a statement of no difference (equality)

Definition 6.5. An *alternative hypothesis* is a statement that stands in opposition to the null hypothesis H_0. It specifies what forms of departure from the null hypothesis are of concern. An alternative hypothesis is denoted as H_a. It is of the form

$$\text{Parameter} \neq \# \quad \text{or} \quad \text{Parameter} > \# \quad \text{or} \quad \text{Parameter} < \#$$

Examples (testing the true mean value):

$$\begin{array}{lll} \mathrm{H}_0 : \mu = \# & \mathrm{H}_0 : \mu = \# & \mathrm{H}_0 : \mu = \# \\ \mathrm{H}_a : \mu \neq \# & \underbrace{\mathrm{H}_a : \mu > \# \quad \mathrm{H}_a : \mu < \#}_{\text{one-sided}} \end{array}$$

two-sided

Often, the alternative hypothesis is based on an investigator's suspicions and/or hopes about the true state of affairs.

The **goal** is to use the data to debunk the null hypothesis in favor of the alternative.

1. Assume H_0.
2. Try to show that, under H_0, the data are preposterous. (using probability)
3. If the data are preposterous, reject H_0 and conclude H_a.

The outcomes of a hypothesis test consists of:

The ultimately decision in favor of

True state of affairs is described by		H_0	H_a
	H_0	OK	Type I. error
	H_a	Type II error	OK

P(Type I error) = α
(same α as from CI's)

this is the probability of rejecting H_0 when H_0 is true.

α is fixed before we look at data.

Example 6.11 (Fair coin). Suppose we toss a coin $n = 25$ times, and the results are denoted by $X_1, X_2, \ldots, X_{25}$. We use 1 to denote the result of a head and 0 to denote the results of a tail. Then $X_1 \sim Binomial(1, \rho)$ where ρ denotes the chance of getting heads, so $\mathrm{E}(X_1) = \rho, \mathrm{Var}(X_1) = \rho(1-\rho)$. Given the result is you got all heads, do you think the coin is fair?

Null hypothesis – H_0: the coin is fair

$H_0: \rho = 0.5$

Alternative hypothesis – $H_A: \rho \neq 0.5$

If H_0 was correct, then $P(\text{results are all heads}) = \left(\frac{1}{2}\right)^{25} < 0.000001$

$\Rightarrow$ I don't think this coin is fair (reject H_0 in favor of H_a)

In the real life, we may have data from many different kinds of distributions! Thus we need a universal framework to deal with these kinds of problems.

We have $n = 25 \geq 25$ iid trials $\Rightarrow$ by CLT we know if $H_0: \rho = 0.5 (= EX_1)$, then

$EX_1 = \rho$

$\mathrm{Var} X_1 = \rho(1-\rho)$

$$\frac{\bar{X} - \rho}{\sqrt{\rho(1-\rho)}/\sqrt{n}} \mathrel{\dot\sim} N(0,1).$$

Then the probability of seeing as "weird or weirder" data is

$P(Z$ bigger than 5 or less than $-5)$ $< .000001$!

This is absurd, so we reject H_0.

We observe $\bar{x} = 1$, so

$$\frac{\bar{x} - 0.5}{\sqrt{0.5(1-0.5)}/\sqrt{25}} = \frac{1 - 0.5}{\sqrt{\frac{0.5(0.5)}{25}}} = 5$$

6.3.1 Significance tests for a mean

Definition 6.6. A *test statistic* is the particular form of numerical data summarization used in a significance test. (in the previous example, the test statistic was $\frac{\bar{x} - 0.5}{\sqrt{0.5(1-0.5)/25}} = 5$)

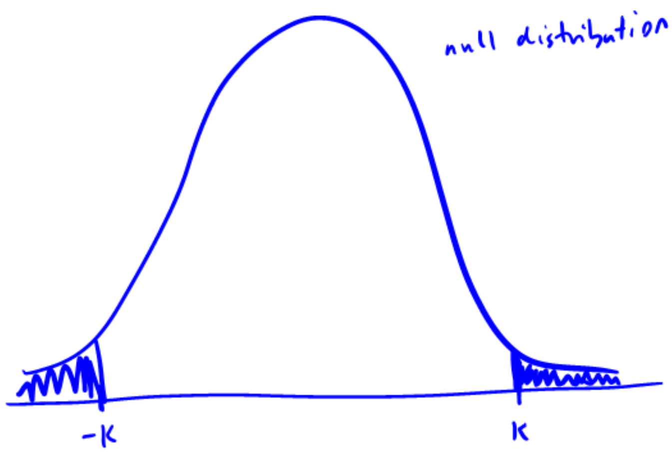
Definition 6.7. A *reference (or null) distribution* for a test statistic is the probability distribution describing the test statistic, provided the null hypothesis is in fact true.

(in the previous example, the null distribution was $N(0,1)$)

Definition 6.8. The *observed level of significance or p-value* in a significance test is the probability that the reference distribution assigns to the set of possible values of the test statistic that are at least as extreme as the one actually observed.

↳ comes from H_a

(in the previous example, the p-value was $< .000\,001$)



Let K be the test statistic value (based on data)

Say $H_0: \mu = \mu_0$

$H_a: \mu \neq \mu_0$

p-value = P(of seeing data as or more "extreme" as K) = $P(Z < -K$ or $Z > K)$

Based on our results from Section 6.2 of the notes, we can develop hypothesis tests for the true mean value of a distribution in various situations, given an iid sample $X_1, \ldots, X_n$ where $\mathrm{H}_0 : \mu = \mu_0$.

Let K be the value of the test statistic, $Z \sim N(0,1)$, and $T \sim t_{n-1}$. Here is a table of p-values that you should use for each set of conditions and choice of H_a.

Situation	K	two-sided $\mathrm{H}_a : \mu \neq \mu_0$	one-sided $\mathrm{H}_a : \mu < \mu_0$	$\mathrm{H}_a : \mu > \mu_0$	
$n \geq 25, \sigma$ known	$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$P(\|Z\| > K)$	$P(Z < K)$	$P(Z > K)$	compare to Normal(0,1)
$n \geq 25, \sigma$ unknown	$\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	$P(\|Z\| > K)$	$P(Z < K)$	$P(Z > K)$	
$n < 25, \sigma$ unknown (data iid $N(\mu, \sigma^2)$)	$\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	$P(\|T\| > K)$	$P(T < K)$	$P(T > K)$	$\leftarrow$ compare to t_{n-1}

Steps to perform a hypothesis test:

1. State H_0 and H_a
2. State α, significance level, usually a small number 0.1, 0.05, 0.01
3. State form of the test statistic, its distribution under the null hypothesis, and all assumption.
4. Calculate the test statistic and p-value
5. Make a decision based on the p-value
   - if the p-value $< \alpha \Rightarrow$ reject H_0, otherwise we fail to reject H_0.
6. Interpret the conclusion using context of problem.

Example 6.12 (Cylinders)**.** The strengths of 40 steel cylinders were measured in MPa. The sample mean strength is 1.2 MPa with a sample standard deviation of 0.5 MPa. At significance level $\alpha = 0.01$, conduct a hypothesis test to determine if the cylinders meet the strength requirement of 0.8 MPa.

1. $H_0: \mu = 0.8$
   $H_a: \mu > 0.8$.

2. $\alpha = 0.01$

3. Since σ unknown, $n = 40 \geq 25$,
   $K = \dfrac{\bar{X} - 0.8}{s/\sqrt{n}}$ is the test statistic

   I assume $X_1, \ldots, X_{40}$ are iid w/ mean μ and variance σ^2
   Then $K \overset{\sim}{\cdot} N(0,1)$ by the CLT under H_0.

4. $K = \dfrac{1.2 - 0.8}{0.5/\sqrt{40}} = 5.06$

   p-value: $P(Z > 5.06) = 1 - P(Z \leq 5.06)$
   $= 1 - \Phi(5.06)$
   $\approx 1 - 1 = 0$

5. Since p-value $\ll \alpha$, I reject H_0 in favor of H_a.

6. There is overwhelming evidence to conclude that the cylinders meet the strength requirement of 0.8 MPa.

Example 6.13 (Concrete beams). 10 concrete beams were each measured for flexural strength (MPa). The data is as follows.

[1] 8.2 8.7 7.8 9.7 7.4 7.8 7.7 11.6 11.3 11.8

The sample mean was 9.2 MPa and the sample variance was 3.0933 MPa. Conduct a hypothesis test to find out if the flexural strength is different from 9.0 MPa.

1. $H_0: \mu = 9 \quad H_a: \mu \neq 9$

2. Choose $\alpha = 0.01$

3. I will use the test statistic $K = \dfrac{\bar{X} - 9}{s/\sqrt{n}}$ (unknown σ)

   and since $n = 0 < 25$, we must assume $X_1, \dots, X_{10} \overset{iid}{\sim} N(\mu, \sigma^2)$

   Then if our assumptions hold, $K \sim t_{n-1} = t_9$ under the null hypothesis.

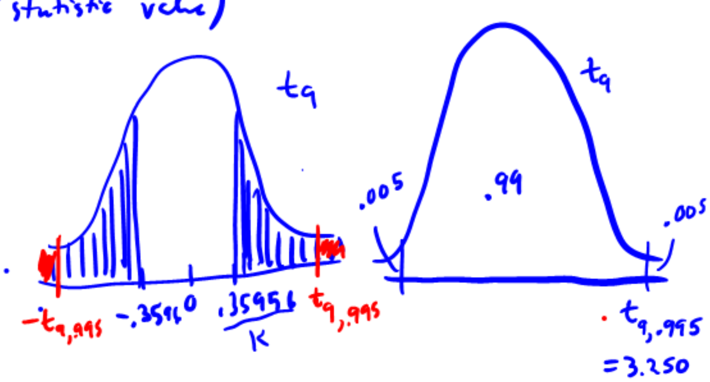
4. $K = \dfrac{9.2 - 9}{\sqrt{\frac{3.0933}{10}}} = 0.3596$

p-value = P(as or more extreme test statistic value)

$T \sim t_9$ $\quad = P(|T| \geq 0.3596)$

$> P(|T| > t_{9,.995})$

$= .01 = \alpha$



5. Since the p-value $> \alpha$, I fail to reject H_0.

6. There is not enough evidence to conclude that the true mean flexural strength of the beams is different from 9 MPa.

6.3.2 Hypothesis testing using the CI

We can also use the $1-\alpha$ confidence interval to perform hypothesis tests (instead of p-values). The confidence interval will contain μ_0 when there is little to no evidence against H_0 and will not contain μ_0 when there is strong evidence against H_0.

Steps to perform a hypothesis test using a confidence interval:

1. State hypotheses H_0 and H_a
2. State the significance level, α
3. State the form of $1-\alpha$ CI along with all assumptions necessary
   - use one-sided CI for one sided tests (i.e. $H_a: \mu < \#$ or $H_a: \mu > \#$) and two-sided CI for two-sided tests ($H_a: \mu \neq \#$)
4. Calculate the CI
5. Based on $1-\alpha$ CI, either reject H_0 (if μ_0 is not in the interval) or fail to reject (if μ_0 is in the interval).
6. Interpret the conclusion in the context of the problem.

Example 6.14 (Breaking strength of wire, cont'd). Suppose you are a manufacturer of construction equipment. You make 0.0125 inch wire rope and need to determine how much weight it can hold before breaking so that you can label it clearly. You have breaking strengths, in kg, for 41 sample wires with sample mean breaking strength 91.85 kg and sample standard deviation 17.6 kg. Using the appropriate 95% confidence interval, conduct a hypothesis test to find out if the true mean breaking strength is above 85 kg.

Example 6.15 (Concrete beams, cont'd)**.** 10 concrete beams were each measured for flexural strength (MPa). The data is as follows.

[1] 8.2 8.7 7.8 9.7 7.4 7.8 7.7 11.6 11.3 11.8

The sample mean was 9.2 MPa and the sample variance was 3.0933 MPa. At $\alpha = 0.01$, test the hypothesis that the true mean flexural strength is 10 MPa using a confidence interval.

Example 6.16 (Paint thickness, cont'd)**.** Consider the following sample of observations on coating thickness for low-viscosity paint.

[1] 0.83 0.88 0.88 1.04 1.09 1.12 1.29 1.31 1.48 1.49 1.59 1.62 1.65 1.71 [15] 1.76 1.83

Using $\alpha = 0.1$, test the hypothesis that the true mean paint thickness is 1.00 mm. Note, the 90% confidence interval for the true mean paint thickness was calculated from before as $(1.201, 1.499)$.

6.4 Inference for matched pairs and two-sample data

An important type of application of confidence interval estimation and significance testing is when we either have *paired data* or *two-sample* data.

6.4.1 Matched pairs

Recall,

Examples:

One simple method of investigating the possibility of a consistent difference between paired data is to

1.

2.

Example 6.17 (Fuel economy)**.** Twelve cars were equipped with radial tires and driven over a test course. Then the same twelve cars (with the same drivers) were equipped with regular belted tires and driven over the same course. After each run, the cars gas economy (in km/l) was measured. Using significance level $\alpha = 0.05$ and the method of critical values, test for a difference in fuel economy between the radial tires and belted tires. Construct a 95% confidence interval for true mean difference due to tire type.

car	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0
radial	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0	7.4	4.9	6.1	5.2
belted	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.7	6.0	4.9

Example 6.18 (End-cut router)**.** Consider the operation of an end-cut router in the manufacture of a company's wood product. Both a leading-edge and a trailing-edge measurement were made on each wooden piece to come off the router. Is the leading-edge measurement different from the trailing-edge measurement for a typical wood piece? Do a hypothesis test at $\alpha = 0.05$ to find out. Make a two-sided 95% confidence interval for the true mean of the difference between the measurements.

piece	1.000	2.000	3.000	4.000	5.000
leading_edge	0.168	0.170	0.165	0.165	0.170
trailing_edge	0.169	0.168	0.168	0.168	0.169

6.4.2 Two-sample data

Paired differences provide inference methods of a special kind for comparison. Methods that can be used to compare two means where two different *unrelated* samples will be discussed next.

Examples:

Notation:

6.4.2.1 Large samples ($n_1 \geq 25, n_2 \geq 25$)

The difference in sample means $\overline{x}_1 - \overline{x}_2$ is a natural statistic to use in comparing μ_1 and μ_2.

If σ_1 and σ_2 are **known**, then Proposition 5.1 tells us

$\mathrm{E}(\overline{X}_1 - \overline{X}_2) =$

$\mathrm{Var}(\overline{X}_1 - \overline{X}_2) =$

If, in addition, n_1 and n_2 are large, So, if we want to test $\mathrm{H}_0 : \mu_1 - \mu_2 = \#$ with some alternative hypothesis, σ_1 and σ_2 are known, and $n_1 \geq 25, n_2 \geq 25$, then we use the statistic

$K =$

which has a $N(0,1)$ distribution if

1. H_0 is true
2. The sample 1 points are iid with mean μ_1 and variance σ_1^2, and the sample 2 points are iid with mean μ_2 and variance σ_2^2.

The confidence intervals (2-sided, 1-sided upper, and 1-sided lower, respectively) for $\mu_1 - \mu_2$ are:

If σ_1 and σ_2 are **unknown**, and $n_1 \geq 25, n_2 \geq 25$, then we use the statistic

$K =$

and confidence intervals (2-sided, 1-sided upper, and 1-sided lower, respectively) for $\mu_1 - \mu_2$:

Example 6.19 (Anchor bolts). An experiment carried out to study various characteristics of anchor bolts resulted in 78 observations on shear strength (kip) of 3/8-in. diameter bolts and 88 observations on strength of 1/2-in. diameter bolts. Let Sample 1 be the 1/2 in diameter bolts and Sample 2 be the 3/8 indiameter bolts. Using a significance level of $\alpha = 0.01$, find out if the 1/2 in bolts are more than 2 kip stronger (in shear strength) than the 3/8 in bolts. Calculate and interpret the appropriate 99% confidence interval to support the analysis.

- $n1 = 88, n2 = 78$
- $\overline{x}_1 = 7.14, \overline{x}_2 = 4.25$
- $s_1 = 1.68, s_2 = 1.3$

6.4.2.2 Small samples

If $n_1 < 25$ or $n_2 < 25$, then we need some **other assumptions** to hold in order to complete inference on two-sample data.

A test statistic to test $\mathrm{H}_0 : \mu_1 - \mu_2 = \#$ against some alternative is $K =$

Also assuming - H_0 is true, - The sample 1 points are iid $N(\mu_1, \sigma_1^2)$, the sample 2 points are iid $N(\mu_2, \sigma_2^2)$, - and the sample 1 points are independent of the sample 2 points.

Then $K \sim$

$1 - \alpha$ confidence intervals (2-sided, 1-sided upper, and 1-sided lower, respectively) for $\mu_1 - \mu_2$ under these assumptions are of the form:

Example 6.20 (Springs)**.** The data of W. Armstrong on spring lifetimes (appearing in the book by Cox and Oakes) not only concern spring longevity at a 950 N/mm^2 stress level but also longevity at a 900 N/mm^2 stress level. Let sample 1 be the 900 N/mm^2 stress group and sample 2 be the 950 N/mm^2 stress group. Let's do a hypothesis test to see if the sample 1 springs lasted significantly longer than the sample 2 springs.

900 N/mm2 Stress	950 N/mm2 Stress
216, 162, 153, 216, 225, 216, 306, 225, 243, 189	225, 171, 198, 189, 189, 135, 162, 135, 117, 162

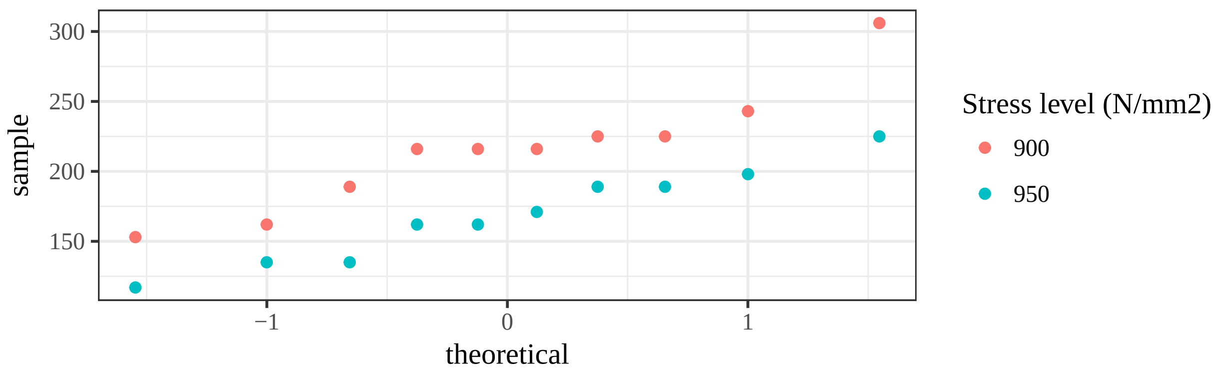


Figure 2: Normal plots of spring lifetimes under two different levels of stress.

Example 6.21 (Stopping distance)**.** Suppose μ_1 and μ_2 are true mean stopping distances (in meters) at 50 mph for cars of a certain type equipped with two different types of breaking systems. Suppose $n_1 = n_2 = 6$, $\overline{x}_1 = 115.7$, $\overline{x}_2 = 129.3$, $s_1 = 5.08$, and $s_2 = 5.38$. Use significance level $\alpha = 0.01$ to test $\mathrm{H}_0 : \mu_1 - \mu_2 = -10$ vs. $\mathrm{H}_A : \mu_1 - \mu_2 < -10$. Construct a 2-sided 99

6.5 Prediction intervals

Methods of confidence interval estimation andd hypothesis testing concern the problem of reasoning from sample information to statements about underlying *parameters* of the data generation (such as μ).

Sometimes it is useful to not make a statement about a parameter value, but create bounds on other *individual values* generated by the process.

> How can we use out data $x_1, \dots, x_n$ to create an interval likely to contain one additional (as yet unobserved) value x_{n+1} from the same data generating mechanism?

Let $X_1, \dots, X_n$ be iid Normal random variables with

$$\mathrm{E}(X_i) = \mu \text{ for all } i = 1, \dots, n$$
$$\mathrm{Var}(X_i) = \sigma^2 \text{ for all } i = 1, \dots, n$$

Then,

Let X_{n+1} be an additional observation from the same data generating mechanism.

$\mathrm{E}(\overline{X}_n - X_{n+1}) =$

$\mathrm{Var}(\overline{X}_n - X_{n+1}) =$

So,

Generally, σ is unknown, so replace σ by s, and

Then, $1-\alpha$ **Prediction intervals** for X_{n+1} are

Table B.4

t Distribution Quantiles

ν	$Q(.9)$	$Q(.95)$	$Q(.975)$	$Q(.99)$	$Q(.995)$	$Q(.999)$	$Q(.9995)$
1	3.078	6.314	12.706	31.821	63.657	318.317	636.607
2	1.886	2.920	4.303	6.965	9.925	22.327	31.598
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.849
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

This table was generated using MINITAB.