

Outline

Chapter 5: Inference for a single population

Introductory Statistics for Engineering Experimentation

Peter R. Nelson, Marie Coffin and Karen A.F. Copeland

Slides by Douglas Bates

5.1 Central Limit Theorem

5.2 A confidence interval for μ

5.3 Prediction and tolerance intervals

5.4 Hypothesis tests

The Central Limit Theorem

- ▶ The central limit theorem is one of the most important results in mathematical statistics. It says that the sample means from a *random sample* (meaning independent samples from a stable process) will be normally distributed, regardless of what the original distribution was, when n is sufficiently large.
- ▶ Formally, if $\mathcal{Y}_1, \mathcal{Y}_2, \dots, \mathcal{Y}_n$ is a random sample from a distribution with $\sigma^2 < \infty$ then for large samples, $\bar{\mathcal{Y}}$ is approximately normally distributed.
- ▶ This is a remarkably powerful result; first, because it is very general and secondly because it is a description of the asymptotic or “limiting” distribution but it holds for quite small values of n .

Other properties of the distribution of the sample mean

- ▶ If the random variables $\mathcal{Y}_1, \mathcal{Y}_2, \dots, \mathcal{Y}_n$ are a random sample (sometimes also described as a “independent and identically distributed” or i.i.d. sample) from a distribution with mean μ and variance σ^2 then $E(\bar{\mathcal{Y}}) = \mu$ and $\text{Var}(\bar{\mathcal{Y}}) = \sigma^2/n$.
- ▶ So the central limit theorem states that, for large n ,

$$\bar{\mathcal{Y}} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- ▶ Exactly how large n must be depends on the form of the original distribution. If it is continuous and reasonably symmetric then $n = 15$ may be large enough. If it is skewed but continuous we may need $n = 30$ or more. For discrete and skewed we may need as much as $n = 100$.
- ▶ Although in practice we only have one sample and one average, \bar{y} we can use computer simulation to consider the sorts of samples we could have gotten and the distribution of the statistic.

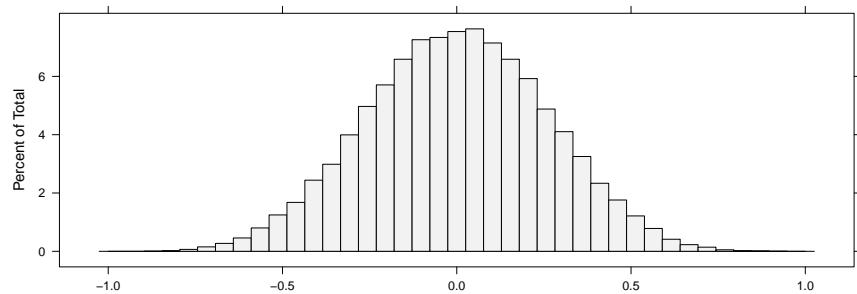
Conducting a simulation study (not part of the course)

- ▶ Suppose we wish to simulate the value of a statistic (e.g. mean or median or variance or standard deviation) from samples of size n drawn from a certain distribution. Let K be the number of replicates we want to obtain.
- ▶ The **sample size**, n , is typically small. The number of replicates, K , can be very large. The larger the value of K , the more accurately we can determine the distribution of the statistic. With modern computers we can afford to use values of K in the hundreds of thousands or more.
- ▶ First determine how to evaluate the statistic from a single sample of size n then use the **replicate** function to repeat the process K times.

Mean of samples of size 5 from $U(-1,1)$

What is the shape of the distribution of the mean of a sample of size $n = 5$ from a $U(-1, 1)$ distribution?

```
> mns5 <- replicate(50000, mean(runif(5, min = -1, max = 1)))
> histogram(~mns5, breaks = seq(-1, 1, len = 40))
```



Sampling densities of statistics

- ▶ The idiom

```
replicate(K, <statfn>(r<distab>(n, <pars>)))
```

produces K replicates of the statistic calculated by **<statfn>** (examples are **mean**, **median**, **var** and **sd**) on samples of size **n** from distribution **<distab>** with parameter(s) **<pars>**.

- ▶ Typically K is large and n is small. Values of 10,000 or 100,000 are used for K on modern computers. The larger the value of K the smoother the approximation to the sampling density. n is the size of the actual sample you can afford to collect.

Effect of changing the sample size, n

- ▶ Performing multiple simulations allows us to see how characteristics of the distribution of \bar{Y} depends on n .

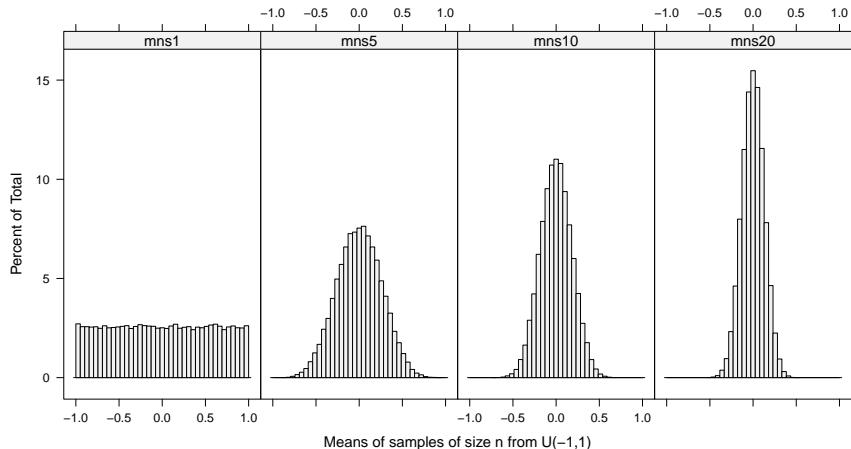
```
> mns1 <- runif(50000, -1, 1)
> mns10 <- replicate(50000, mean(runif(10, -1, 1)))
> mns20 <- replicate(50000, mean(runif(20, -1, 1)))
> sapply(list(mns1, mns5, mns10, mns20), mean)
[1] -0.0017200996  0.0005271466 -0.0008944196 -0.0006052555
```

```
> sapply(list(mns1, mns5, mns10, mns20), var)
[1] 0.33399089 0.06681132 0.03306059 0.01671302
```

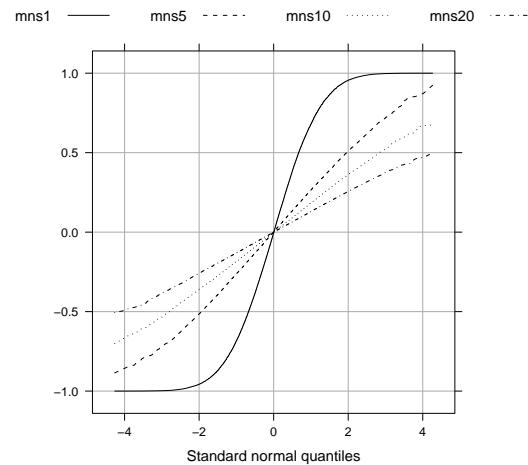
- ▶ As n increases the expected value of the sample mean stays near 0.
- ▶ As n increases the variance of the sample mean decreases. Roughly, $V(\bar{X}_n) = \frac{1}{3} \cdot \frac{1}{n}$

Shape of distribution of \bar{X}_n

- As n increases, the shape of the distribution of \bar{X}_n tends to the “bell-curve” or Gaussian shape and it has less variability. That is, it tends to a “central limit”.



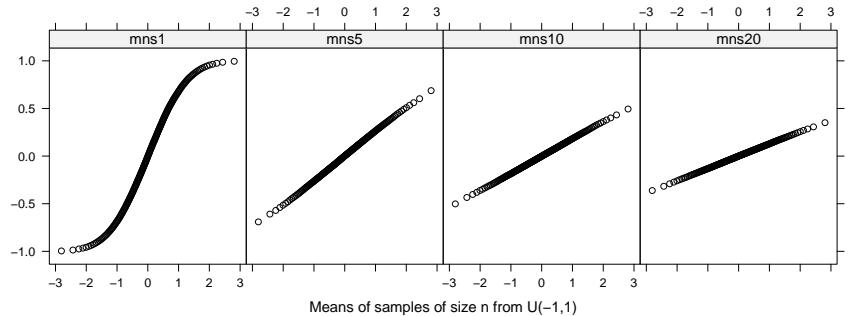
Overlaid normal probability plots for \bar{Y}_n



The conclusion is that the distribution of means from an i.i.d. sample of a uniform distribution is very close to a normal, even for $n = 5$.

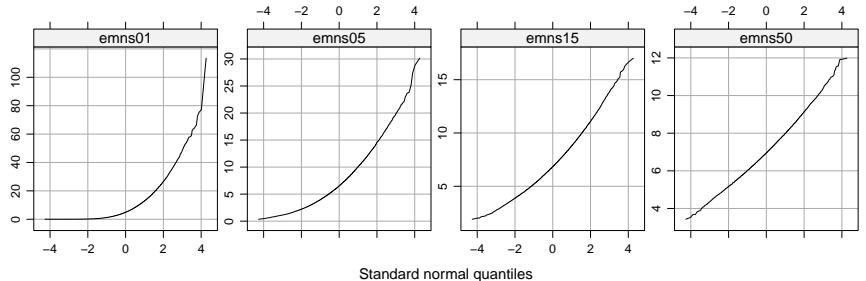
More detail on the shape of the distribution of \bar{Y}

- In addition to the histogram we can use normal probability plots to evaluate the deviations of the distribution of \bar{Y} from normality.



Sample means from an exponential distribution

```
> emns01 <- replicate(50000, mean(rexp(1, rate = 1/7)))
> emns05 <- replicate(50000, mean(rexp(5, rate = 1/7)))
> emns15 <- replicate(50000, mean(rexp(15, rate = 1/7)))
> emns50 <- replicate(50000, mean(rexp(50, rate = 1/7)))
```



Even for $n = 50$ there is noticeable skewness in the distribution (although we would not be far wrong in assuming normality at $n = 50$).

Elementary uses of the C.L.T.

- If we have plausible values of the variance of our process, perhaps from a pilot study, we can use the normal distribution and the Central Limit Theorem (C.L.T.) to evaluate probabilities regarding the sample mean.
- Example 5.1.3 discusses product lifetimes that have an unknown mean and a variance of approximately 8 years. The number of products to sample so that we are 95% certain that \bar{y} will be within 1 year of the true mean is derived from

$$0.95 = P(|\bar{Y} - \mu| < 1)$$

The distribution of \bar{Y} will be approximately normal with mean μ and standard deviation σ/\sqrt{n} . For a standard normal, 95% of the probability is within “2” standard deviations of the mean (the actual multiple is `qnorm(0.025) = -1.95996`) so we want $1 = qnorm(0.025)^2 \frac{8}{n}$. That is, $n >$

```
> 8 * qnorm(0.025)^2
```

```
[1] 30.73167
```

Confidence intervals

- Our “best guess” at a parameter is called a *point estimate*. For example, we usually use the sample mean, \bar{y} , as the point estimate of μ .
- An *interval estimate* or *confidence interval* is an interval of plausible values for the parameter. Values outside the interval are “unreasonable” and values inside are “not unreasonable”.
- To calibrate the meaning of “unreasonable” we assign a value α to the probability of getting data like we did or even more extreme when the parameter is outside. This corresponds to the “p-value” in a hypothesis test.
- The *coverage probability* or *confidence level* is $1 - \alpha$. Typically we set $\alpha = 0.05$ or $\alpha = 0.01$ resulting in 95% or 99% confidence intervals.
- Formally, the coverage probability is the probability that an interval constructed in this way will cover the true parameter value.

Approximations for binomial or Poisson distributions

- The text describes approximations of the probabilities for a binomial or Poisson distribution based on the normal distribution.
- These are interesting from the point of view of understanding that these distributions will tend to have a “bell-curve” shape when n is large and p is moderate for the binomial or λt is large for the Poisson.
- In practice, though, you can evaluate probabilities for such distributions exactly so there is no need to use approximations.

A confidence interval on μ

- In the unlikely event that someone were to tell us what the standard deviation, σ , of the population was but somehow not know much about the mean, μ , we could create a $(1 - \alpha)$ confidence interval as

$$\bar{y} \pm z(\alpha/2) \frac{\sigma}{\sqrt{n}}$$

where $z(\alpha/2)$ is the **upper** $\alpha/2$ quantile of the standard normal distribution.

- For example, the upper 0.025 quantile of the standard normal is

```
> qnorm(0.025, low = FALSE)
```

`[1] 1.959964`
so a 95% confidence interval on μ for this artificial, “known sigma” case is

$$\bar{y} \pm 1.959964 \frac{\sigma}{\sqrt{n}}$$

Use of Student's T distribution

- In the real world no one tells us what σ is and we must estimate it as s . A statistician named William Gossett, who wrote under the pseudonym "A Student", derived the distribution of the shifted, scaled sample mean when the scale is based on the estimate, s , not the theoretical value σ .
- This distribution is called the "Student's t distribution". It is similar to the standard normal distribution but a bit more spread out. The spreading depends on the number of "degrees of freedom" in the estimate of σ^2 . The degrees of freedom are written as ν . For a single sample $\nu = n - 1$.
- As ν increases the T distribution approaches the standard normal. If we were using tables we would call anything with $\nu > 30$ a standard normal. When using a computer we don't bother.
- Notation: the t distribution with ν degrees of freedom is written $t(\nu)$. The corresponding R functions are `dt`, `pt`, `qt` and `rt`. The upper α quantile is written $t(\alpha; \nu)$.

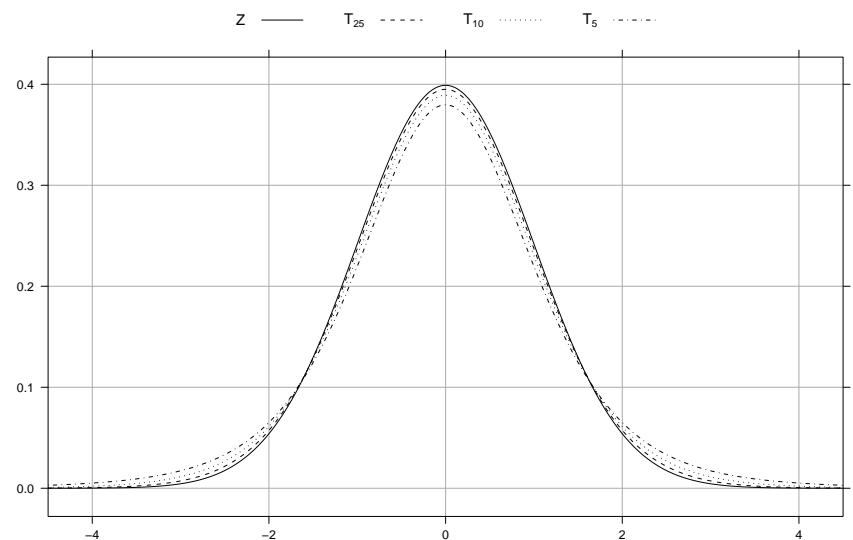
General form of the confidence interval

- The general form of the confidence interval on μ is

$$\bar{y} \pm t\left(\frac{\alpha}{2}, n - 1\right) \frac{s}{\sqrt{n}}$$

- We can use this formula for any values of n . If n is large we don't need strong assumptions on the shape of the original distribution. If n is small we must assume that the original distribution is close to the normal (but, of course, we can't check this with a small sample - a "Catch 22" situation).
- The R function to create this interval is `t.test`. The name comes from the corresponding hypothesis test, which we will discuss later.

Graphical comparison of $t(\nu)$ and Z



Example 5.2.2

The example provides (probably fictitious) discharge times for a particular electric vehicle

```
> sd(charge <- c(5.11, 2.1, 4.27, 5.04, 4.47, 3.73, 5.96, 6.21))
```

```
[1] 1.3108
```

```
> summary(charge)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.10	4.14	4.76	4.61	5.32	6.21

```
> t.test(charge)
```

```
One Sample t-test
```

```
data: charge
```

```
t = 9.9502, df = 7, p-value = 2.211e-05
```

```
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
```

```
3.5154 5.7071
```

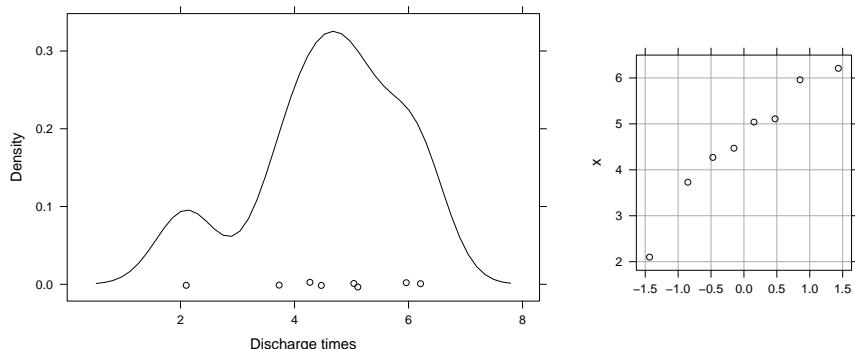
```
sample estimates:
```

```
mean of x
```

```
4.6113
```

Example 5.2.2 (cont'd)

Because the degrees of freedom, $\nu = 7$, are quite small we should check for normality.



Another *R* evaluation of confidence intervals

- ▶ Another way of evaluating a confidence intervals on μ is with the `confint` function, which provides confidence intervals on the parameters in a fitted model.
- ▶ To use `confint` we fit what we sometimes call the “trivial” model

$$\mathcal{Y}_i = \mu + \epsilon_i, \quad i = 1, \dots, n$$

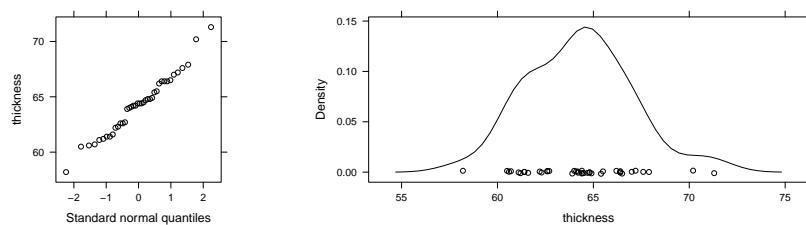
The estimate of μ , $\hat{\mu} = \bar{y}$ will be called (`Intercept`) in the output. The formula for the model contains the constant term, `1`, as the only predictor.

```
> confint(fm1 <- lm(charge ~ 1))
```

2.5 % 97.5 %

(Intercept) 3.5154 5.7071

Clear-coat thickness (example 5.2.4)



```
> with(ccthickn, summary(thickness))
Min. 1st Qu. Median Mean 3rd Qu. Max.
58.2   62.3   64.4   64.3   66.2   71.3
> sd(ccthickn$thickness)
[1] 2.7176
> confint(fm2 <- lm(thickness ~ 1, ccthickn))
2.5 % 97.5 %
(Intercept) 63.391 65.129
```

Clear-coat thickness (cont'd)

The summary of the fit of the “trivial” model includes many of the statistics from the data.

```
> summary(fm2)
Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.2600    0.4297 149.5 <2e-16
Residual standard error: 2.718 on 39 degrees of freedom
```

The values in this summary include

$\bar{y} = 64.26$ The parameter estimate, $\hat{\mu}$.

$s = 2.718$ The sample standard deviation, $\hat{\sigma}$

$n - 1 = 39$ The degrees of freedom, ν , for the variance estimate, s^2 .

$\frac{s}{\sqrt{n}} = 0.4297$ The standard error of the mean, $\sqrt{\text{Var}(\mathcal{Y})}$

Sample sizes

- The half-width of a confidence interval, also called the *margin of error* depends on
 - The **confidence level** Higher confidence levels require wider intervals
 - The **standard deviation** More variability in the original distribution results in wider intervals.
 - The **sample size** Larger samples produce narrower intervals.
- Given a working value for σ we can determine the sample size needed to attain a given margin of error.
- If we are willing to assume that n is large we can use $z(\alpha/2)$ in the calculation. For small n it gets tricky because $\nu = n - 1$ determines the multiplier which, in turn, affects the sample size. We must solve a nonlinear equation but computers are good at that.

Sample size calculations

- Example 5.2.5 shows calculations for the sample size from the formula $n = \left[\frac{t(\alpha/2; \infty)s}{d} \right]^2$ when the desired margin of error, d , is 0.2, the working value of s is 0.4 and α is 5% and we round the answer to the next largest integer.
`> ceiling(qnorm(0.025)*0.4/0.2)^2`

[1] 16

- Because this is a small value of n we should instead solve for n in $n = \left[\frac{t(\alpha/2; n-1)s}{d} \right]^2$

```
> ceiling(uniroot(function(x) x-(qt(.025,x-1)*0.4/0.2)^2,  
+ c(2,100))$root)
```

[1] 18

Section 5.3: Prediction and tolerance intervals

- A confidence interval on μ provides a measure of the precision of the information regarding the unknown population parameter. It does not directly tell us about bounds on where we expect a future observation to fall.
- A *prediction interval* indicates where a single future observation is likely to be.
- A *tolerance interval* indicates where a large proportion of the population is likely to be.
- Unlike the confidence interval on μ , prediction intervals and tolerance intervals depend strongly on the shape of the distribution of the data.
- In theory one can make a confidence interval arbitrarily narrow by taking a sufficiently large sample. You can't do this for a prediction interval.

Prediction intervals on a future observation

- If it is reasonable to assume that the *data* (i.e. $\mathcal{Y}_1, \mathcal{Y}_2, \dots, \mathcal{Y}_n$) are from normal distribution then we could say that a model for the data is

$$\mathcal{Y}_i = \mu + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Our estimate $\hat{\mu} = \bar{\mathcal{Y}}_n$ is independent of ϵ_{n+1} . The variability in the difference between \mathcal{Y}_{n+1} and $\bar{\mathcal{Y}}_n$ is the sum of the variability in $\bar{\mathcal{Y}}_n - \mu$ ($\frac{\sigma^2}{n}$) and the variability in ϵ_{n+1} (σ^2).
- Because we estimate σ^2 the $(1 - \alpha)$ prediction interval becomes

$$\bar{y} \pm t(\alpha/2; n-1)s\sqrt{1 + \frac{1}{n}}$$

Evaluating a prediction interval

- ▶ The prediction interval could be evaluated according to the formula. For the clear-coat thickness data the 95% prediction interval on a future thickness measurement is

```
> with(ccchickn, mean(thickness) + c(-1,1) * qt(0.975, 39) *  
+      sd(thickness) * sqrt(1 + 1/40))  
  
[1] 58.69478 69.82522
```

- ▶ An alternative is to use the `predict` function applied to the trivial model and with the optional argument `interval = "pred"`. This produces a matrix with n rows that are identical so I just look at the first row.

```
> predict(fm2, int = "pred")[1,]  
  
   fit      lwr      upr  
64.26000 58.69478 69.82522
```

Tolerance intervals

- ▶ A tolerance interval is more difficult to describe and to calculate than is a prediction interval.
- ▶ Methods for tolerance intervals are given in the text but we will not cover this topic in this course.

Section 5.4 Hypothesis tests

- ▶ A hypothesis test is a procedure for deciding if a particular value of a parameter is reasonable, given the observed data.
- ▶ We have a probability model (e.g. our sample is a random sample from a normal distribution with mean μ and variance σ^2), the observed data, y_1, y_2, \dots, y_n and a particular value of the parameter in mind (e.g. the mean clear-coat thickness should be 65 microns).
- ▶ We consider two competing claims called the *null hypothesis*, written H_0 , and the *alternative hypothesis*, written H_a .
- ▶ H_0 is the “no change” assumption. For our example, it is $\mu = 65$. H_a is the result we are trying to establish. It is also the result indicated by the data. In our example $\bar{y} = 64.26$ microns. If we are interested only in whether we are “on target” then $H_a : \mu \neq 65$. If we are interested in whether the clear coats are systematically too thin then $H_a : \mu < 65$.
- ▶ These are called “two-tailed” and “one-tailed” alternatives, respectively.