Methods of Mathematical Statistics

Notes by Tim Brown and Davide Ferrari

Module 7: Interval Estimation

Contents

1 Confidence Intervals for Means - 7.1

1.1 Probability Reminders

Reminders

Results results from Module 5, Section 4 recalled.

Suppose $X_1, \ldots, X_n$ are a random sample from $N(\mu, \sigma^2)$.

Sample mean: $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$.

Sums of iid normal's squared: If $Z_i = (X_i - \mu)/\sigma$, so that $Z_i \sim N(0,1)$ are independent, then $W = Z_1^2 + \cdots + Z_n^2 \sim \chi^2(n)$, the chisquare distribution with n degrees of freedom.

Sample variance: If $S^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ then

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

$S^2, \bar{X}$ are independent.

Reminders Ctd

Linear combinations: If $X_i \sim N(\mu_i, \sigma_i^2)$ are independent $i = 1, \ldots, n$ then for constants $a_1, \cdots, a_n$

$$Y = \sum_{i=1}^{n} a_i X_i \sim N(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2).$$

t random variable: If $Z \sim N(0,1)$ and $U \sim \chi^2(r)$ are independent then

$$T = \frac{Z}{\sqrt{U/r}} \sim t(r),$$

the t-distribution with r degrees of freedom.

1st use of t distribution is for sample mean and variance: $t = \frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$, but it was also applicable to regression slope and intercept for normal samples or errors.

1.2 Confidence Intervals - known σ

Interval Estimation

How close is the estimator to the parameter?

As the estimator is a random variable we can only make probability statements.

Suppose $X_1, \cdots X_n$ i.i.d. $N(\mu, \sigma^2)$ where σ^2 is known and μ is the parameter an unknown number (frequentist inference).

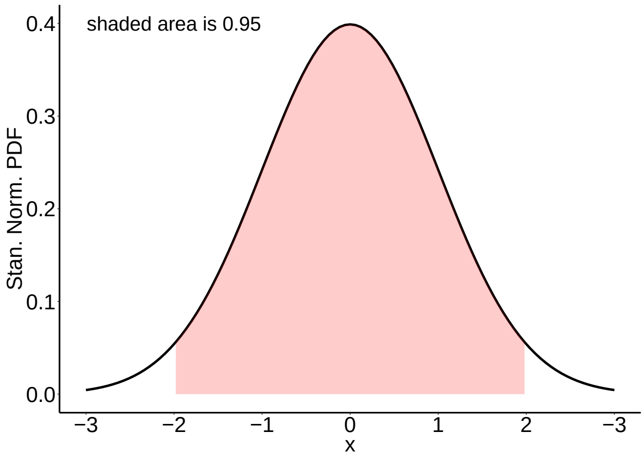

Figure 1: Illustration of $z_{0.025} \approx 2$ with 95% probability shaded pink

Know that $\bar{X} \sim N(\mu, \sigma^2/n)$ is the mle. Thus, for $0 < \alpha < 1$, $z_{\alpha/2}$ can be found so that

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

Figure 1 illustrates this for $\alpha = 0.025$, the answer is $\pi_{0.975}$, the 97.5 percentile of the normal distribution.

Interval Estimation

Rearranging yields

$$P\left\{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha.$$

Or there is probability $1 - \alpha$ that the random interval

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

contains μ.

With data $\bar{x}$, the interval either contains μ or it doesn't.

If α is small, it would be unlucky that the interval did *not* contain μ, because this only happens 100α % of the time.

So we say the interval is a $100(1 - \alpha)\%$ *confidence interval* for the unknown population mean μ.

1.3 Example - Known σ

Example - Confidence Interval Known σ

Suppose $X \sim N(\mu, 1296)$ represents the lifetime of a light bulb (in hours).
Test 27 bulbs, $\bar{x} = 1479$.
Assume population standard deviation is $\sigma = 36$.
A 95% confidence interval for μ is

$$\begin{aligned}&\left[\bar{x} - z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{x} + z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right)\right] \\ &= \left[1478 - 1.96\left(\frac{36}{\sqrt{27}}\right), 1478 + 1.96\left(\frac{36}{\sqrt{27}}\right)\right] \\ &= [1464, 1492].\end{aligned}$$

In other words, we are 95% confident that [1464,1492] contains the true value of the population mean for lightbulb life.

1.4 Width of Confidence Interval & Sample Size

Width of Confidence Interval

The smaller σ the shorter the interval.

The shorter the interval, the more reliable is our estimate.

Can also decrease the width by increasing sample size, if this is feasible.

Will use simulation to examine the interpretation in the computer labs.

1.5 Example - Non-normal distribution, Large Sample Size

Example - Orange Juice

If distribution is not normal, we can use the central limit theorem: if n is large enough, $(\bar{X} - \mu)/(\sigma/\sqrt{n}) \approx N(0, 1)$.

Eg: X is the amount of orange juice consumed (grams per day) by an Australian. Know $\sigma = 96$.

Sampled 576 Australians and found $\bar{x} = 133$ grams per day.

An approximate 90% confidence interval for the mean amount of orange juice consumed by an Australian, regardless of the underlying distribution for individual orange juice consumption, is

$$133 \pm 1.645\left(\frac{96}{\sqrt{576}}\right) = [126, 140] \text{(nearest integer)}$$

Figure 2 illustrates the choice of the multiplier 1.645.

Often n is not large in science, because observations can be expensive (eg clinical or agricultural trials).

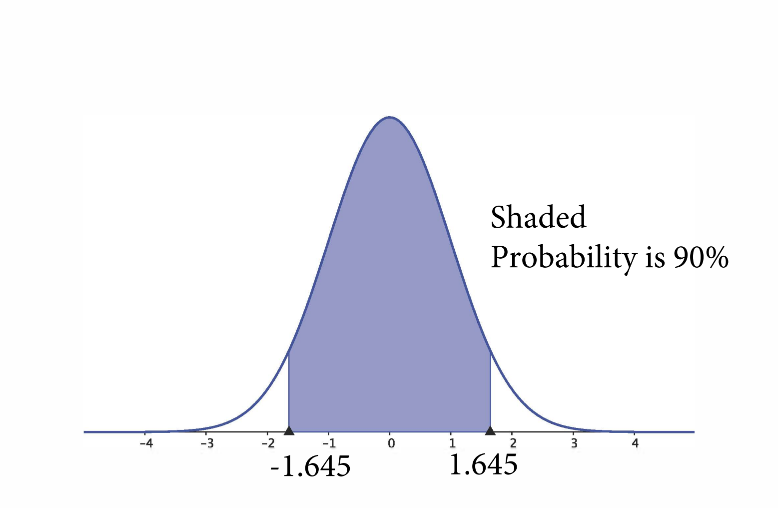

Figure 2: Multiplier for normal 90% confidence interval

1.6 Pivots - Definition

Pivots

A random variable $Q(X_1, \cdots, X_n; \theta)$ is a *pivot* if its distribution is independent of unknown parameters – that is, $Q(X_1, \cdots, X_n; \theta)$ has the same distribution for all values of θ.

Have seen this for normal data with known variance $(\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$ so $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ is a pivot.

Then for any set $\mathcal{A}$, $P\{Q(X_1, \cdots, X_n; \theta) \in \mathcal{A}\}$ does not depend on θ.

For example, in the normal case with known variance,

$$P\left(a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b\right)$$

does not depend on μ.

Knowing distributions

Allows us to construct confidence intervals if we know the distribution of $Q(X_1, \cdots, X_n; \theta)$.

So can use logic to construct the confidence interval if the pivot can be found.

Rearrange *pivot* probability statements to make unknown parameter the subject and data values at the ends.

1.7 Example - exponential

Lightbulb life exponential

Assume instead that the lightbulbs had a lifetime which is exponentially distributed with mean μ.
A confidence interval can be found using the Gamma distribution.
The pivotal quantity is $\frac{\sum_{i=1}^n X_i}{\mu} \sim \text{Gamma}(n, 1)$ if $X_1, \cdots, X_n$ are independent random variables with exponential distribution mean μ.
So

$$P(\gamma_{\alpha/2} \leq \frac{\sum_{i=1}^n X_i}{\mu} \leq \gamma_{1-\alpha/2}) = 1 - \alpha,$$

where $\gamma_{\alpha/2}, \gamma_{1-\alpha/2}$ are the $\alpha/2, 1-\alpha/2$ quantiles of the Gamma distribution with shape parameter n and scale parameter 1.
Equivalently,

$$P(\frac{\sum_{i=1}^n X_i}{\gamma_{1-\alpha/2}} \leq \mu \leq \frac{\sum_{i=1}^n X_i}{\gamma_{\alpha/2}}) = 1 - \alpha.$$

Lightbulb life exponential

Thus, a $100(1-\alpha)$ percent confidence interval is $\left[\frac{\sum_{i=1}^n X_i}{\gamma_{1-\alpha/2}}, \frac{\sum_{i=1}^n X_i}{\gamma_{\alpha/2}}\right]$. The R commands show the calculations:

```
gq <- qgamma(c(0.025, 0.975), scale = 1, shape = 27)
# mean is 1479 so sum is 1479*27
c(27 * 1479/gq[2], 27 * 1479/gq[1])

## [1] 1048.220 2244.288
```

Note the contrasting result to the assumption of normality with a standard deviation of 36. Difference is that the standard deviation for an exponential distribution is the same as the mean, so its estimate is 1479. The normal sd reflects planned obsolescence and the exponential is a memoryless lightbulb.

1.8 One sample t-confidence interval - unknown σ

One sample t-confidence interval

Suppose $X_1, \cdots X_n$ i.i.d. $N(\mu, \sigma^2)$ where σ^2 is also unknown.

Know that

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$$

where t_{n-1} is the t distribution with $n-1$ degrees of freedom

Now proceed as before.

The reason this works is that both $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ and $\frac{\sqrt{n}(\bar{X}-\mu)}{S}$ are pivots.

One sample t-confidence interval

For α in $(0,1)$ choose $t_{\alpha/2}(n-1)$ so that

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

Rearranging yields

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

And for observed $\bar{x}$ and s, a $100(1-\alpha)\%$ confidence interval for μ is

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

1.9 Example - one sample t confidence interval

Example - One sample t CI

Suppose $X \sim N(\mu, \sigma^2)$ is the amount of butterfat produced (in pounds) by a cow.

Sample of $n = 20$ cows and observed $\bar{x} = 507.50$ and $s = 89.75$.

Now $t_{0.05}(19) = 1.729$ (see Figure 3) so a 90% confidence interval for μ is

$$507.50 \pm 1.729\left(\frac{89.75}{\sqrt{20}}\right) = [472.8, 542.20]$$

R output given below - hypothesis test will be covered in Module 8.

```
Butterfat <- c(481, 537, 513, 583, 453, 510, 570,
    500, 457, 555, 618, 327, 350, 643, 499, 421, 505,
    637, 599, 392)
t.test(Butterfat, conf.level = 0.9)$conf.int

## [1] 472.7982 542.2018
## attr(,"conf.level")
## [1] 0.9

sd(Butterfat)

## [1] 89.75082
```

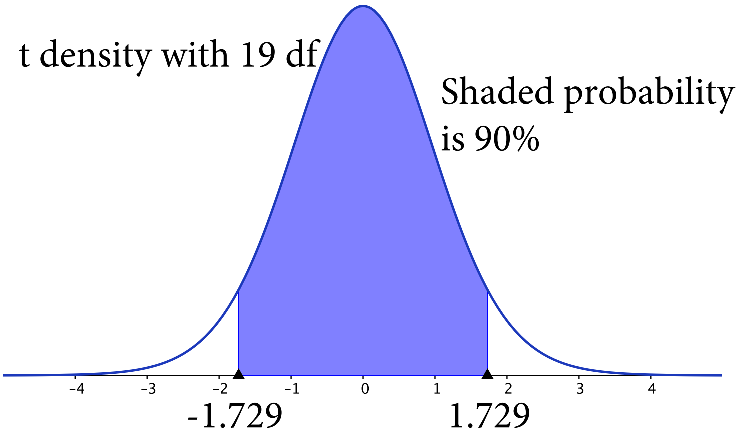

Figure 3: t density with 19 df - 90% probability shaded

Comment - One sample t

The t- distribution is appropriate if sample is from a normal population.

Check using QQplot above - some evidence of departure from normality but not great evidence against normality.

If not normal and n large can construct approximate confidence intervals using the normal distribution with sample sd.

OK if distribution is continuous, symmetric and unimodal for moderate n.

If not normal and small sample size, distribution free methods for the median can be used - see 7.5.

t- confidence intervals, of level $100(1-\alpha)\%$ using the t-distribution are always of the form:

$$\text{estimate} \pm t_{\alpha} \times \text{estimated standard error} \quad (1)$$

recalling that the estimated standard error is an estimate of the standard deviation of the estimate (eg $\frac{S}{\sqrt{n}}$ for $\frac{S}{\sqrt{n}} = sd(\bar{X})$.)

Lightbulb example using normal approximation

The exponential lightbulb example could also be analysed approximately by assuming 27 is large enough that $\bar{X} \approx N(\mu, \frac{\mu^2}{n})$ where μ is the population mean lifetime. This leads to the confidence interval shown in the R output:

```
1479 - qnorm(0.975) * 1479/sqrt(27)

## [1] 921.1282
```

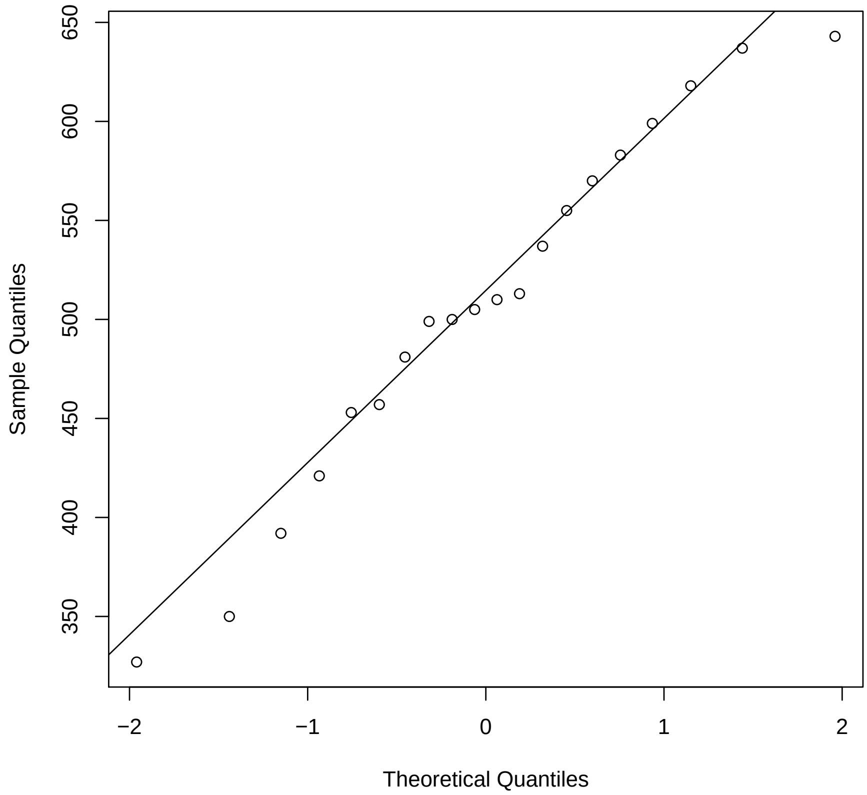

Figure 4: Normal Quantile Plot to check Normality - probably OK

```r
1479 + qnorm(0.975) * 1479/sqrt(27)

## [1] 2036.872
```

This is interval is similar to the exact one calculated using the pivot showing the approximation is reasonably close but not very close.

1.10 One sided confidence intervals

One-sided confidence intervals

One sided probability statements about the pivot can give one-sided confidence intervals:

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

Yields

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

And $[\bar{x} - z_\alpha\left(\sigma/\sqrt{n}\right), \infty]$ is a one sided confidence interval for μ (Gives a lower bound for known σ).

Or $[\bar{x} - t_\alpha(n-1)\left(S/\sqrt{n}\right), \infty]$ with unknown σ .

1.11 Example - One sided confidence intervals

Example - One sided confidence interval

A winemaker requires a minimum concentration of 10g per litre of sugar in the grapes used to make a certain wine. In a sample of 30 units he finds an average concentration of 11.9 grams per litre with standard deviation 0.96.

Figure 5 illustrates that $t_{0.05}(29) = 1.6991$

So a 95% one-sided confidence interval comes from

$$\bar{x} - t(29)_\alpha\left(\frac{\sigma}{\sqrt{n}}\right) = 11.9 - 1.6991\frac{0.96}{\sqrt{30}} = 11.60$$

and thus the winemaker is 95% confident the average sugar content is above 11.61 grams per litre.

2 Confidence Intervals for Difference of Two Means - 7.2

2.1 Difference of Two Means - σ known

Difference 2 Means - σ known

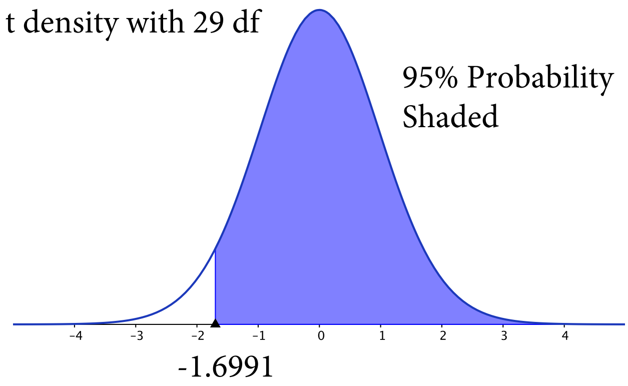

Figure 5: 5th percentile for t distribution with 29 df is -1.6991

Supppose there are two population means, μ_X, μ_Y, with interest centring on the difference.

Have independent samples, $X_1, \cdots, X_n$ i.i.d. $N(\mu_X, \sigma_X^2)$, $Y_1, \cdots, Y_m$ i.i.d. $N(\mu_Y, \sigma_Y^2)$,

Assume σ_X^2 and σ_Y^2 are known. Then

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim N(0, 1).$$

is the pivot.

Difference 2 Means - σ known

Hence

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

And rearranging as usual gives the $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ as

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sigma_w$$

where $\sigma_w = \sqrt{\sigma_X^2/n + \sigma_Y^2/m}$.

Rare to know the population variances!

2.2 Difference of Two Means - σ unknown

Difference 2 Means - σ unknown, samples large

If σ_X^2 and σ_Y^2 not known and n and m are large

Can replace σ_X and σ_Y by sample sd's S_X and S_Y

Obtain *approximate* confidence intervals using the central limit theorem

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

CLT says distribution of the pivot $\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{S_X^2/n+S_Y^2/m}} \approx N(0,1)$

Diff. 2 Means - σ unknown, common variance

If the sample size is small and σ_X^2 and σ_Y^2 are not known:

Assume: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ so that a pivot may be found

Know

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma^2/n + \sigma^2/m}} \sim N(0,1)$$

As the samples are independent,

$$U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

as U is the sum of independent χ^2 random variables

Further U and Z are independent

Diff. 2 Means - σ unknown, common variance

From the definition of a T random variable,

$$T = \frac{Z}{\sqrt{U/(n+m-2)}} \sim t_{n+m-2}. \qquad (2)$$

which is thus the pivot

Some algebra shows

$$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_P\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Where

$$S_P = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}$$

is the pooled estimate of the common variance - note that the unknown σ cancels

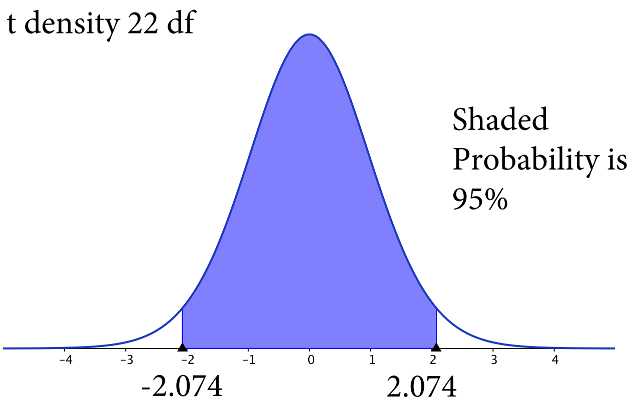

Figure 6: 95% Prob for t(22) between -2.074 and 2.074

Diff. 2 Means - σ unknown, common variance

Can find t_0 so that

$$P(-t_0 \leq T \leq t_0) = 1 - \alpha$$

And rearranging as usual gives a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ as

$$\bar{x} - \bar{y} \pm t_0 s_P \sqrt{\frac{1}{n} + \frac{1}{m}}$$

Where

$$s_P = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}}$$

Example - Independent Groups Test Scores

Suppose two independent groups take the same test. Assume the scores are normally distributed and have a common unknown population variance

Further suppose $n = 9$, $m = 15$, $\bar{x} = 81.31$, $\bar{y} = 78.61$, $s_x^2 = 60.76$, $s_y^2 = 48.24$.

Pivot from (2) has df $9 + 15 - 2 = 22$, and Figure 6 $t_{0.025}(22) = 2.074$ so 95% confidence interval is

$$81.31 - 78.61 \pm 2.074\sqrt{\frac{8(60.76) + 14(48.24)}{22}}\sqrt{\frac{1}{9} + \frac{1}{15}}$$
$$= [-3.65, 9.05]$$

Difference 2 Means - σ unknown, different variances

Unequal variances, $\sigma_X^2 \neq \sigma_Y^2$, and m and n small?

Use

$$W = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}}$$

which has an approximate t distribution with complicated degrees of freedom, r (Welch 1949). See p 367 of the text.

Most computer packages can compute the degrees of freedom

Often the default for constructing confidence intervals

Example - Independent Groups 2 means

Example. Force required to pull wires apart two types of wire, X and Y- 20 repititions. Find a 95% confidence interval for the population mean force required.

```
X <- c( 28.8, 24.4, 30.1, 25.6, 26.4, 23.9 22.1, 22.5,
,27.6, 28.1, 20.8, 27.7, 24.4, 25.1, 24.6,
26.3, 28.2, 22.2, 26.3, 24.4)
Y <- c(14.1, 12.2, 14.0, 14.6,  8.5, 12.6, 13.7, 14.8, 14.1, 13.2, 12.1, 11.4, 10.1, 14.2,
13.1, 11.9, 14.8, 11.1, 13.5)
boxplot(Wires$X,Wires$Y, names=c("X","Y"))
```

Boxplot in Figure 7 shows different means and possibly different variances for the two types.

```
t.test(X, Y, conf.level = 0.95)$conf.int

## [1] 11.23214 13.95786
## attr(,"conf.level")
## [1] 0.95

t.test(X, Y, conf.level = 0.95, var.equal = T)$conf.int

## [1] 11.23879 13.95121
## attr(,"conf.level")
## [1] 0.95
```

Welch approximate t-distribution appropriate so 95% confidence interval is [11.23,13.96]

If equal variances assumed confidence interval is narrower but might be too narrow - [11.24,13.95].

Not a big difference!

t values illustrated for 38 and 33 df in Figures 8 and 9 .

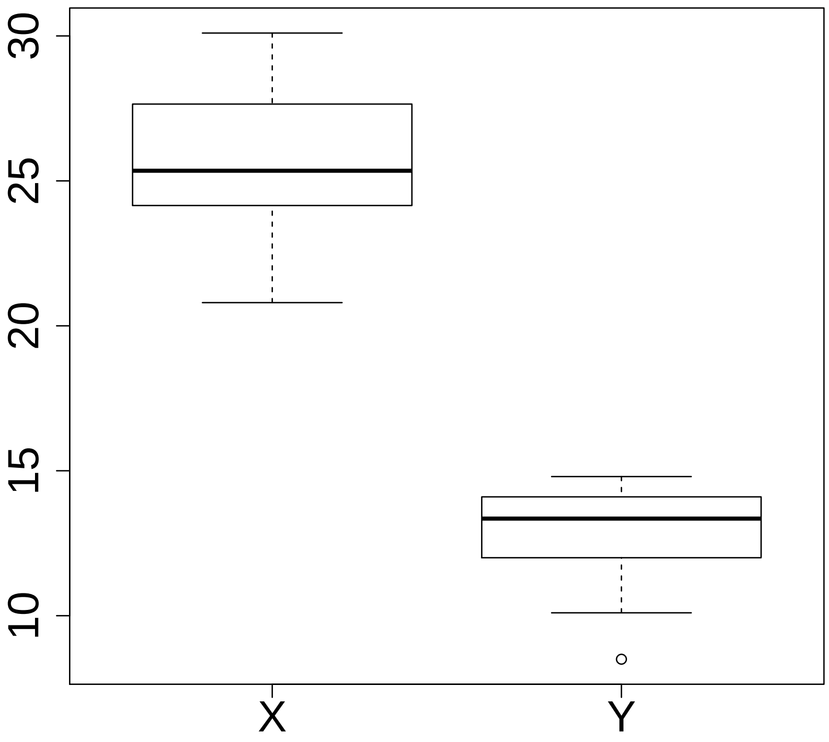

Figure 7: Box Plots of Forces for X and Y

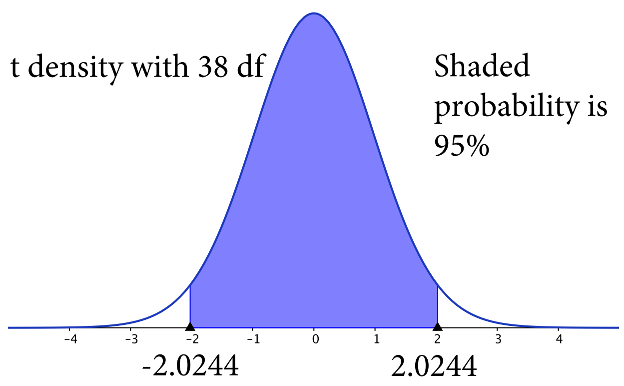

Figure 8: t 38 with 95% shaded

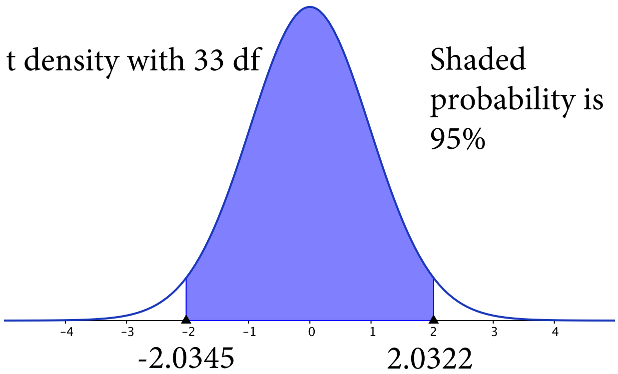

Figure 9: t 38 with 95% shaded

2.3 Paired data - t confidence intervals

Paired t-confidence intervals

Again interested in difference between means of two sets of observations, $\mu_D = \mu_X - \mu_Y$

Observe n independent pairs of rv's $(X_1, Y_1), \cdots, (X_n, Y_n)$ but generally each X is dependent on the Y

Let $D_i = X_i - Y_i$

Often reasonable to suppose $D_i \sim N(\mu_D, \sigma_D^2)$

CI's for $\mu_D = \mu_X - \mu_Y$ now come from t confidence intervals for the single sample of $D's$

$100(1-\alpha)\%$ confidence interval for μ_D is

$$\bar{d} \pm t_{\alpha/2}(n-1)\frac{s_d}{\sqrt{n}}$$

Example - paired t-confidence intervals

The reaction times (in seconds) to a red or green light for 8 people are given in the following table. Find a 95% confidence interval the difference in reaction time for the mean difference across all people.

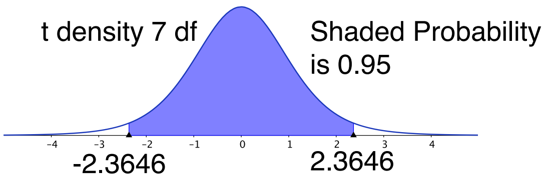

Figure 10: Traffic Lights Example

	Red(X)	Green(Y)	D=X-Y
1	0.30	0.24	0.06
2	0.43	0.27	0.16
3	0.23	0.36	-0.13
4	0.32	0.41	-0.09
5	0.41	0.38	0.03
6	0.58	0.38	0.20
7	0.53	0.51	0.02
8	0.46	0.61	-0.15

$\bar{d} = -0.0625$, $s_d = 0.0765$, $df = 8 - 1 = 7$, $t_{0.025}(7) = 2.3646$,

$$-0.0625 \pm 2.3646\frac{0.0765}{\sqrt{8}} = [-0.1265, 0.0015] \text{ is the 95\% CI}$$

3 Confidence Intervals for Proportions - 7.3

3.1 Single Sample

Parameter is p - single sample

Observe n Bernoulli trials with unknown probability p of success.

Want a confidence interval for p.

Recall (p.15 in Module 6) that the sample proportion of successes $\hat{p} = \bar{X}$ (where $X_i, i = 1, \cdots, n$ are the 0-1 rv's that are 1 at each success) is the maximum likelihood estimator for p and is unbiased for p.

The central limit theorem shows for large n,

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx N(0, 1).$$

Rearranging as usual and replacing p by $\hat{p}$ (they are close because the sample size is assumed large) in the variance yields the approximate $100(1-\alpha)\%$ confidence interval as

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Interval Estimation

An alternative approach does not use the estimator of p in the denominator. The p's in the confidence interval satisfy

$$\frac{|\,\hat{p}-p\,|}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}$$

.

This is the same as

$$H(p) = (\hat{p}-p)^2 - \frac{z_{\alpha/2}^2 p(1-p)}{n} \leq 0$$

Let $z_0 = z_{\alpha/2}$. To find these values of p, note that H is a quadratic in p which has a minimum and has zeros at

$$\frac{\hat{p} + z_0^2/(2n) \pm z_0\sqrt{\hat{p}(1-\hat{p})/n + z_0^2/(4n^2)}}{1+z_0^2/n}.$$

These zero's are the endpoints of the confidence interval and, for large n, the approaches give similoar answers.

3.2 Example - Newspoll - single sample proportion CI

Example - Newspoll: single sample proportion

In the Newspoll of 3rd April 2017 (see p.31 of Module 3) 36% of 1708 voters sampled said they would vote for the Government first if an election were held on that day. What is a 95% confidence interval for the population proportion of voters who would vote for the Government first?

The sample proportion has an approximate normal distribution since the sample size is large so the required confidence interval is

$$0.36 \pm 1.96\sqrt{\frac{0.36 \times 0.64}{1708}} = [0.337, 0.383]$$

noting that $1.96 = z_{0.025}$ (see Figure 1) .

So it would be unlucky if the true proportion was greater than 38% or less than 34% - 37% is in the middle of this range so appears plausible.

3.3 Example: Newspoll - Bayesian probability interval

Example: Newspoll - Bayesian probability interval

Calculate a 95% probability intervaln for p if a uniform prior distribution, h, on (0,1) is assumed.

No. of voters saying they would vote for the Government first, $Y \sim Bin(1708, p)$, given unknown probability $0 < p < 1$.

Got $y = 0.36 \times 1708 = 615$ and so $1708 - 615 = 1093$ would vote for another party first.

Use Module 6 equation (21) to calculate the posterior density, $k(p|Y = 615)$.

Example: Newspoll - Bayesian prob. int. Ctd

Posterior density is given, for $0 < p < 1$, by

$$\begin{aligned} k(p|Y = 615) &= \frac{h(p)P(Y = 615|p)}{\int_{-\infty}^{\infty} P(Y = 615|u)h(u)\, du} \\ &= \frac{P(Y = 615|p)}{\int_0^1 P(Y = 615|u)h(u)\, du} \\ &= \frac{p^{615}(1-p)^{1093}}{\int_0^1 u^{615}(1-u)^{1093}\, du} \end{aligned}$$

since the combinatorial factors $\binom{1708}{615}$ cancel

Fact: for any numbers $\alpha > 0, \beta > 0$

$$\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\, du = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{3}$$

and this is often called the *Beta* function $B(\alpha, \beta)$

Example: Newspoll - Bayesian prob. int. Ctd 2

The posterior density is called the Beta$(615+1, 1093+1)$ density and for general $\alpha > 0, \beta > 0$, the *Beta*(α, β) density, f is 0 outside (0,1) and for $0 < x < 1$ satisfies

$$f(x) = x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha, \beta) \tag{4}$$

Mean is easy to calculate using equation (4) and $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$

$$\begin{aligned} \text{posterior mean} &= \int_0^1 p \times p^{615}(1-p)^{1093}/B(616, 1094)\, dp \\ &= B(617, 1094)/B(616, 1094) \\ &= \frac{\Gamma(617)\Gamma(1094)\Gamma(1710)}{\Gamma(616)\Gamma(1094)\Gamma(1711)} = \frac{616}{1710} \end{aligned} \tag{5}$$

Example: Newspoll - Bayesian prob. int. Ctd 3

Little difference between the MLE = 615/1708 and the mean of the posterior distribution 616/1710 (0.0002!)

To get a probablity interval containing 95% probability, it is good first to look at the posterior density, which was plotted in Mathematica - see Figure 11 and this was blown up to examine [0.32,0.4] in Figure 12

Commands were

```
x <- seq(from = 0, to = 1, by = 0.001)
y <- dbeta(x, shape1 = 616, shape2 = 1094)
plot(x, y, type = "l", ylab = "posterior pdf(x)")
plot(x[0.32 < x & x < 0.4], y[0.32 < x & x < 0.4],
    type = "l", ylab = "posterior pdf(x)", xlab = "x")
```

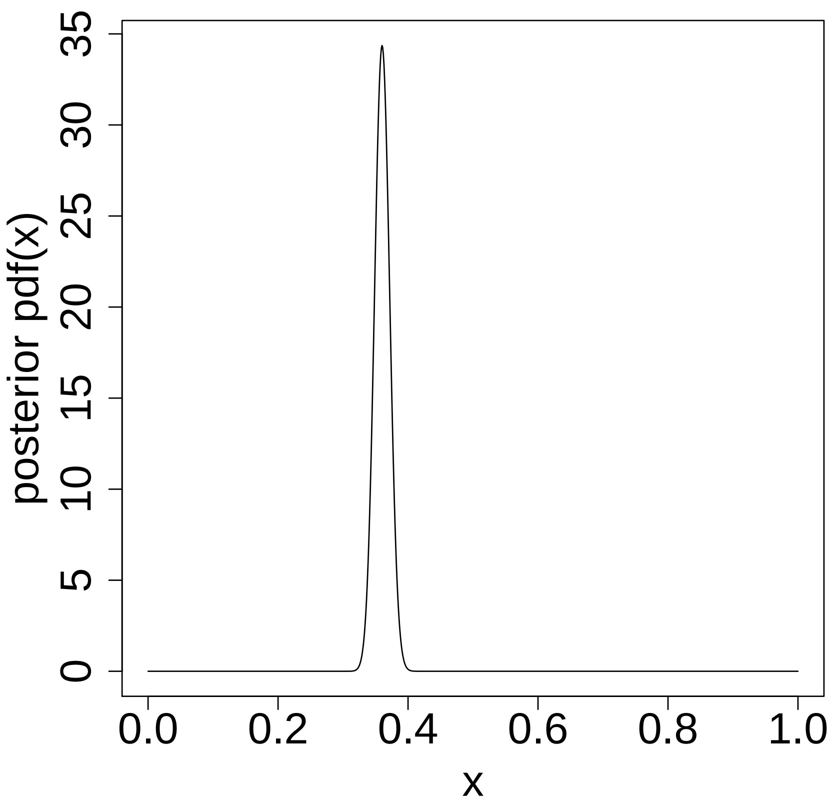

Figure 11: Posterior Probability Density Function - Shape? - [0.32,0.4]?

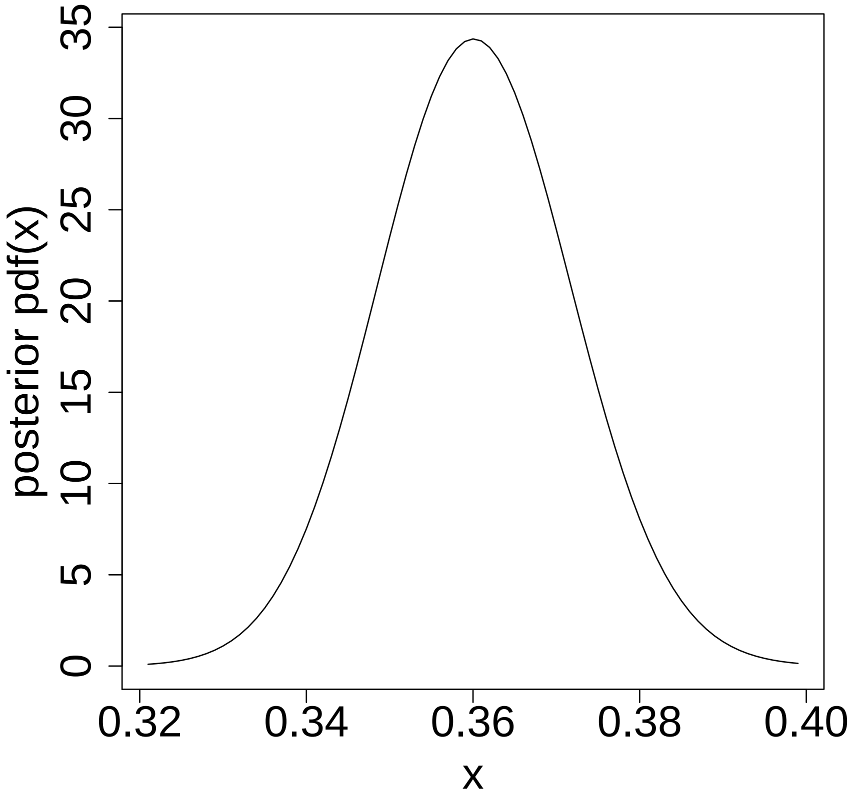

Figure 12: Posterior Probability Density Function on Interval [0.3,0.4]

Example: Newspoll - Bayesian prob. int. Ctd

The simplest way to find a 95% posterior probability interval for p is to use the 2.5% and 97.5% quantiles of the posterior distribution.

The R commands and output are:

```
qbeta(c(0.025, 0.975), shape1 = 616, shape2 = 1094)

## [1] 0.3376449 0.3831327
```

Since the posterior distribution is evidently close to normal, an approximation would be the normal quantiles.

The variance of the beta α, β distribution is $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$.

So the R commands to get an approximate 95% posterior probability interval are:

```
616/1710 + c(-1, 1) * 1.96 * sqrt(616 * 1094/1710^2/1711)

## [1] 0.3374864 0.3829814
```

Example: Newspoll - Summary

The Bayesian and frequentist approaches give very similar answers - all are close the confidence interval using the normal approximation.

This if often true for large sample sizes where the data dominates the choice of method, as long as these are sensible like maximum likelihood estimation or Bayesian methods.

For small sample sizes very different results can be obtained.

3.4 Sufficient Statistics - 6.7

Sufficiency definition

The order of the successes in a sequence of Bernoulli trials is information that might be considered, but it does not give any extra information about p over the total number of successes in the trials.

This is because, if the number of successes, m, in n trials is known, the $\binom{n}{m}$ posssibilities have equal probabilities that do not depend on p.

In any estimation problem based on observations $X_1, \cdots X_n$ for a parameter θ, a statistic T is *sufficent* for θ, if the joint distribution of the observations $X_1, \cdots X_n$, conditional on the value of T, does not depend on θ.

Once T is known, there is no more information about θin the sample.

So just use functions of T as estimators for θ.

Sufficiency of the number of trials

An equivalent definition for sufficiency is that the joint pmf or pdf *factorises* into a function that depends on the value of T and θ together with one which is a function of the data values alone, $f(x_1, \cdots, x_n) = g(\theta, T(x_1, \cdots, x_n))h(x_1, \cdots, x_n)$.

Good example is that the *number of successes* is a *sufficient* statistic for the success probability p, so estimation and confidence intervals should be based only on the *number of successes*.

MLE is proportion of successes and this is the sufficient statistic divided by the number of trials - MLE is always a function of the sufficient statistic.

Bayesian inference is the same using the conditional distribution for the whole sample given the parameter as using the conditional distribution of the sufficient statistic.

3.5 Two proportions

Two proportions

Suppose there are two sets of Bernoulli trials with different probabilities of success p_1, p_2 and numbers of trials n_1, n_2 and that all the trials are independent.

The sample proportions of successes in the two trials are $\hat{p}_1 = Y_1/n_1, \hat{p}_2 = Y_2/n_2$ where Y_1, Y_2 are the numbers of successes and are the sufficient statistics for p_1, p_2.

Further for $i = 1, 2$ and large n_i, $Y_i \sim \text{Bin}(n_i, p_i)$, so $E(Y_i/n_i) = p_i$, $Var(Y_i/n_i) = p_i(1-p_i)/n_i$, independence and the CLT gives:

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \approx N(0,1).$$

On substituting $\hat{p}_i$ for p_i in the variance, $100(1-\alpha)\%$ CI is:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Example - Difference between successive Newspolls

At the previous poll, with 1824 voters sampled, there were 37% of voters who reported that they would vote for the Government first. What is a 90% confidence interval for the difference in proportions in the population on the two occasions?

CI is

$$\begin{aligned}&0.36 - 0.37 \pm 1.6449\sqrt{\frac{0.36 \times 0.64}{1708} + \frac{0.37 \times 0.63}{1824}} \\ &= [-0.037, 0.0.017]\end{aligned}$$

So with 90% confidence the difference contains 0, corresponding to no change in public opinion.

Note: unlike the previous analysis this allows for sampling variability in both polls, so this would be preferred in analysing the Australian headline that the vote had dropped.

Interval Estimation - Summary

Confidence intervals are straightforward to construct if we know or can approximate the sampling distribution of the statistic and can construct a pivot.

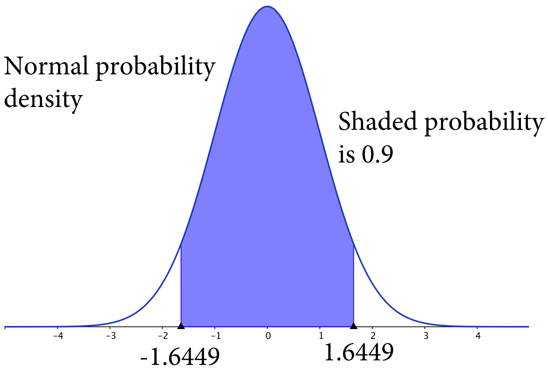

Figure 13: Multiplier for Difference in Proportions

We have looked at some well known (and widely used) examples for means and proportions.

Tricky interpretation for Frequentist confidence interval as the proportion of times in repeated sampling that the interval contains the true parameter.

More straightforward for Bayes because it *is* an interval that has posterior probability of 95% of containing the parameter given the data.

95% is the conventional level in science because they are a convenient way to report results of an experiment.

4 Sample Size - 7.4

4.1 Sample Size: Means

Sample Size: Means

Often a researcher requires a degree of precision in their results.

This is measured by the width of the condfidence interval.

Example: $X_1, \cdots X_n$ i.i.d $N(\mu, 15^2)$. Want a 95% confidence interval of width 2. (i.e. $\bar{x} \pm 1$).

Confidence interval given by $\bar{x} \pm 1.96\frac{15}{\sqrt{n}}$.

So we need

$$1.96\frac{15}{\sqrt{n}} = 1,$$

which yields

$$\sqrt{n} = 29.4, \quad \text{or} \quad n \approx 864.36$$

so a sample size of 865 or more will give a confidence interval width of 1.

Sample Size: Means

Simple confidence interval has the form:

$$\bar{x} \pm \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \bar{x} \pm \epsilon.$$

For a given ϵ need

$$\epsilon = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \text{ or } n = \frac{z_{\alpha/2}^2\sigma^2}{\epsilon^2}.$$

Example: Means

A researcher plans to select a sample of first-grade girls in order to estimate the mean height μ. Sample is required to be large enough so that a researcher is 95% confident the sample mean will be within 0.5 cm of μ. From previous studies knows $\sigma \approx 2.8$cm.

$$n = \frac{z_{\alpha/2}^2\sigma^2}{\epsilon^2} = \frac{1.96^2(2.8^2)}{0.5^2} = 120.47$$

121 girls or more will meet the requirement for the width of the confidence interval.

4.2 Sample Size:Proportions

Sample Size: Proportions

Confidence interval is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

So $\hat{p} \pm \epsilon$ gives

$$\epsilon = z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

Or

$$n = \frac{z_{\alpha/2}^2\hat{p}(1-\hat{p})}{\epsilon^2}$$

Can use preliminary estimate of $\hat{p}$ if this is available.

Otherwise note $\hat{p}(1-\hat{p}) \leq 1/4$ so $n = \frac{z_{\alpha/2}^2}{4\epsilon^2}$ is conservative.

Example: Proportions

Rate of unemployment rate has been 8%. Take new sample and want to be 99% sure the new estimate is within 0.001 of true proportion.

$$n = \frac{z_{\alpha/2}^2\hat{p}(1-\hat{p})}{\epsilon^2} = \frac{2.576^2(0.08)(1-0.08)}{0.001^2} \approx 488,394$$

At which stage the researcher panics and says I don't really need to be that sure.

98% confidence and a difference of 0.01 gives $n = 3,982$ which is more realistic.

5 Distribution Free Confidence Intervals for Percentiles - 7.5

5.1 Order Statistics used to give Confidence Intervals

Order Statistics

Sometimes assumptions of normality of the distribution or large sample size are not true.

Instead *distribution free* confidence intervals are designed to be valid for all distributions, perhaps only requiring that the underlying distribution is continuous.

Here we consider consider confidence intervals for percentiles.

For example, we will obtain a confidence interval for the median say, without making many assumptions about the underlying distribution.

To do this we consider order statistics. They have many applications.

Order Statistics

$X_1, \cdots, X_n$ i.i.d.

$$\begin{aligned} Y_1 &= \text{smallest of the } X_i \\ Y_2 &= \text{2nd smallest of the } X_i \\ &\vdots \\ Y_n &= \text{largest of the } X_i \end{aligned}$$

Order statistics

$$Y_1 < Y_2 < \cdots < Y_n$$

Often see notation $X_{(i)} = Y_i$ in the literature.

Order Statistics

$Y_1 < \cdots < Y_5$ order statistics associated with a random sample $X_1, \ldots, X_5$ with p.d.f. $f(x) = 2x$, $0 < x < 1$.

Consider $P(Y_4 < 1/2)$. This occurs if at least four of the X_i are less than 1/2.

Have 5 Bernoulli trials with probability of success given by

$$P(X_i < \frac{1}{2}) = \int_0^{1/2} 2x dx = \left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

$$\begin{aligned} P(Y_4 < \frac{1}{2}) &= P(\text{at least 4 } X_i\text{'s} < \frac{1}{2}) \\ &= P(\text{exactly 4 } X_i\text{'s} < \frac{1}{2}) + P(\text{exactly 5 } X_i\text{'s} < \frac{1}{2}) \end{aligned}$$

Order Statistics

which is

$$P(Y_4 < \frac{1}{2}) = \binom{5}{4} 0.25^4 0.75 + 0.25^5 = 0.0156.$$

Now

$$F(x) = \int_0^x 2t dt = t^2 \big|_0^x = x^2.$$

Order Statistics

Thus in general

$$G(y) = P(Y_4 < y) = \binom{5}{4} (y^2)^4 (1 - y^2) + (y^2)^5$$

so that after taking derivatives the p.d.f. is

$$g(y) = G'(y) = \frac{5!}{3!1!} (y^2)^3 (1 - y^2)(2y) = \frac{5!}{3!1!} F(y)^3 \{1 - F(y)\} f(y)$$

as we have seen that $F(x) = x^2$.

Order Statistics

$Y_1 < Y_2 < \cdots < Y_n$ order statistics for a sample from a continuous distribution with cdf $F(x)$ and pdf $f(x) = F'(x)$.

$$\begin{aligned} G_r(y) &= P(Y_r \leq y) \\ &= \sum_{k=r}^{n} \binom{n}{k} F(y)^k (1 - F(y))^{n-k} \\ &= \sum_{k=r}^{n-1} \binom{n}{k} F(y)^k (1 - F(y))^{n-k} + F(y)^n \end{aligned}$$

With some patience, the pdf can be obtained. However, there is an informal argument which gives the correct answer quickly.

Order Statistics

For small values of dy, the pdf, g_r of Y_r satisfies

$$g_r(y)dy \approx P(y < Y_r \leq y + dy).$$

For $y < Y_r \leq y + dy$ need one X_i in $(y, y + dy]$, $r - 1 \leq y$ and $n - r > y + dy$. The probability is the same whichever X_i's are chosen to be in the three intervals, $(-\infty, y], (y, y + dy], (y + dy, \infty)$, so consider one choice $B = [X_1, \cdots, X_{r-1} < y, X_r \in (y, y + dy], X_{r+1}, \cdots, X_n > y + dy]$. Because $P(X_r \in (y, y + dy]) \approx f(y)dy, P(X_{r+1} > y + dy) \approx 1 - F(y)$,

$$P(B) \approx F(y)^{r-1} f(y) dy (1 - F(y))^{n-r}.$$

There are $n\binom{n-1}{r-1}$ ways to choose the rv's to go into the three intervals.

Order Statistics

Hence, taking the limit as $dy \to 0$, we have the density of the rth order statistic

$$g_r(y) = \frac{n!}{(r-1)!(n-r)!} F(y)^{r-1}\{1 - F(y)\}^{n-r} f(y)$$

pdf of smallest order statistic is

$$g_1(y) = n(1 - F(y))^{n-1} f(y)$$

and the pdf of the largest is

$$g_n(y) = nF(y)^{n-1} f(y).$$

For a uniform distribution, the order statistics have a Beta distribution, and this demonstrates the value of the Beta distribution constant (taking $\alpha = r, \beta = n + 1 - r$.)

Order Statistics

$X_1, \dots, X_4$ from uniform $[0, \theta]$ distribution. $Y_1, \dots, Y_4$ are the order statistics.

Likelihood is

$$L(\theta) = \left(\frac{1}{\theta}\right)^4, \quad 0 \le x_i \le \theta,\ i = 1, \dots, 4$$

and is zero if $\theta < x_i$ for some i.

This is maximised when θ is as small as possible, so $\hat{\theta} = \max(X_i) = Y_4$

Now,

$$g_4(y_4) = \frac{4!}{3!1!} \left(\frac{y_4}{\theta}\right)^3 \left(\frac{1}{\theta}\right) = 4\frac{y_4^3}{\theta^4}$$

Order Statistics

Then

$$E(Y_4) = \int_0^\theta y_4 4 \frac{y_4^3}{\theta^4} dy_4 = \frac{4}{5}\theta$$

so the maximum likelihood estimator Y_4 is biased.

But $(5/4)Y_4$ is unbiased.

Can further show for $0 < c < 1$,

$$1 - c^4 = P(c\theta < Y_4 < \theta) = P(Y_4 < \theta < Y_4/c)$$

so a $100(1 - c^4)\%$ confidence interval for θ is $[y_4, y_4/c]$.

If $c = 0.05^{1/4} = 0.47$ this gives a 95% confidence interval from y_4 to $2.11y_4$.

Percentiles

Recall for a continuous distribution, $F(X) \sim U(0,1)$.

To see this, take $0 \leq w \leq 1$,

$$P(F(X) \leq w) = P(X \leq F^{-1}(w)) = F(F^{-1}(w)) = w,$$

so the density is 1 for $0 \leq w \leq 1$ and $F(X) \sim U(0,1)$.

Moreover, $F(Y_1) < F(Y_2) < \cdots < F(Y_n)$ (as F is nondecreasing).

Percentiles

So $W_i = F(Y_i)$ are order statistics from $U(0,1)$ distribution.

The pdf of rth order statistic $W_r = F(Y_r)$ is

$$h_r(w) = \frac{n!}{(r-1)!(n-r)!} w^{r-1}\{1-w\}^{n-r}.$$

And hence can obtain, from the argument for the Beta distribution, $E(W_r) = r/(n+1)$, $r = 1, \cdots n$.

Percentiles

$100p$th percentile π_p has probability p to the left of π_p, so $\pi_p = F^{-1}(p)$.

Since $E(F(Y_r)) = r/(n+1)$, $F(Y_r)$ is an unbiased estimator of $r/(n+1)$ for every F.

So it makes sense to use $Y_r = F^{-1}(F(Y_r))$ to estimate $\pi_p = F^{-1}(p)$ where $p = r/(n+1)$.

And this is the reason that $100p$th sample percentile is taken as Y_r where $r = (n+1)p$.

If this is not an integer, we take a weighted average of the adjacent order statistics.

Percentiles

For example the sample median is

$$\tilde{m} = \begin{cases} Y_{(n+1)/2} & \text{when } n \text{ is odd} \\ \frac{Y_{n/2}+Y_{(n/2)+1}}{2} & \text{when } n \text{ is even} \end{cases}$$

Confidence Intervals for Percentiles

Can use sample percentiles to estimate distribution percentiles.

How precise?

Or what are the corresponding confidence intervals?

CI Percentiles - n = 5

Order statistics $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ for iid rv's $X_1, \cdots, X_5$

Then Y_3 is an estimator of the median $m = \pi_{0.5}$.

For the true median to be between Y_1 and Y_5 must have at least one $X_i < m$ but not five $X_i < m$.

If the distribution of the $X's$ is continuous, $P(X < m) = 0.5$.

And if W is the number of the $X's < m$, then $W \sim Bin(5, 0.5)$ and

$$\begin{aligned} P(Y_1 < m < Y_5) &= P(0 < W < 5) \\ &= \sum_{k=1}^{4} \binom{5}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{5-k} \\ &= 1 - 0.5^5 - 0.5^5 = \frac{15}{16} \approx 0.94 \end{aligned}$$

So (y_1, y_5) is a 94% confidence interval for m.

CIs for Percentiles

In general, want i and j so that, to the closest possible extent,

$$\begin{aligned} P(Y_i < m < Y_j) &= P(i - 1 < W < j) \\ &= \sum_{k=i}^{j-1} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} \approx 1 - \alpha \end{aligned} \tag{6}$$

Need to use computed binomial probabilities (R or Mathematica) to determine i and j

Or use the normal approximation to the binomial.

Note that these confidence intervals do not arise from pivots and cannot achieve 95% confidence exactly!

Example - CI Percentiles 9 Fish

Lengths of nine fish in cm: 32.5, 27.6, 29.3,30.1, 15.5, 21.7, 22.8, 21.2, 19.0.

Order statistics are: 15.5, 19.0, 21.2, 21.7, 22.8, 27.6, 29.3, 30.1, 32.5.

And

$$P(Y_2 < m < Y_8) = \sum_{k=2}^{7} \binom{9}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{9-k} = 0.9609375.$$

R command is

```r
pbinom(7, size = 9, prob = 0.5) - pbinom(1, size = 9,
    prob = 0.5)

## [1] 0.9609375
```

So a 96.1% confidence interval for m is $[19.0, 30.1]$.

Confidence Intervals for Percentiles - General

Argument can be extended to any percentile and any order statistics, for example the ith and jth.

If $W \sim Bin(n, \pi_p)$ is the number of the X_i's $< \pi_p$, then

$$\begin{aligned} 1-\alpha &= P(Y_i < \pi_p < Y_j) \\ &= P(i-1 < W < j) \\ &= \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned} \tag{7}$$

Example - CI for Percentiles

Text 6.10-2 $n = 27$ incomes (\$100's). 161, 169, 171, 174, 179, 180, 183, 184, 186, 187, 192, 193, 196, 200,204 205, 213, 221, 222, 229, 241, 243,256 264, 291, 317, 376

25th percentile of the $X-$ distribution is $\pi_{0.25}$.

And the number, W, of the $X's$ which are below $\pi_{0.25}$ satisfies $W \sim Bin(27, 0.25)$.

W $\approx N(\text{mean}27/4 = 6.75, \text{variance} = 81/16))so$

$$P(Y_4 < \pi_{0.25} < Y_{10}) = P(3.5 < W < 9.5)$$
$$\approx \quad \Phi\left(\frac{9.5 - 6.75}{9/4}\right) - \Phi\left(\frac{3.5 - 6.75}{9/4}\right) = 0.815$$

So (\$17,400,\$18,700) is an approximate 81.5% C.I. for the 25th percentile.

The exact value comes from the R output:

```r
pbinom(9, size = 27, prob = 0.25) - pbinom(3, size = 27,
    prob = 0.25)

## [1] 0.8201453
```

showing that the approximation is reasonable (and better with the continuity correction).
The exact calculation would normally be done since R is readily available.