11. Interval Estimation

Readings: Rosner: 6.5-6.11; SAS Lab 4

A) Confidence intervals
B) CI for μ when σ^2 is known
C) CI for μ when σ^2 is unknown: the t-Distribution
D) CI for population variance, σ^2
E) CI for binomial proportion, p
F) One-sided CI

Review:

1. Through repeated random sampling from the population of interest, the sample mean $\overline{X}$ and sample variance s^2 will vary from sample to sample.
2. If the individual observations in a sample, $X_1, X_2, \ldots, X_n$ are *iid* N [μ , σ^2] then

$$\left.\begin{array}{l} \overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right) \\ \dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \end{array}\right\} \text{ independent}$$

3. Even if $X_1, X_2, \ldots, X_n$ are not normally distributed the CLT tells us that, for a reasonably large sample size, $\overline{X}$ is approximately distributed N(μ , σ^2/n)
4. No matter what the distribution of the X_i or the sample size, we have:

$$E\left[\overline{X}\right] = \mu$$
$$V\left[\overline{X}\right] = \frac{\sigma^2}{n}$$
$$E\left[s^2\right] = \sigma^2$$
$$V\left[s^2\right] = \frac{2\sigma^4}{n-1}$$

A) Confidence intervals (CI)

We estimate the population mean μ, a fixed but *unknown* value (parameter) by $\overline{X}$ (a sample statistic). How accurate is a specific value of the estimate? Can we find an interval we're "fairly confident" contains the *true* value of μ?

e.g. An experiment was done to determine the blood alcohol level (mg/ml) required to cause respiratory failure in rats. For 7 rats, the required levels were 9.0, 9.7, 9.4, 9.3, 9.2, 8.9, 9.0. The estimate of the average amount of alcohol required for rats is $\bar{x}$ = 9.21 mg/ml.

Assuming that σ = 0.3 mg/ml), a 95% CI for μ, the true mean alcohol level at failure, is given by (8.95 mg/ml, 9.43 mg/ml). What does this mean? How do these results compare with another study that claimed it took an average of 10 mg/ml to cause failure? How do they compare with a study that claimed it took 9 mg/ml to cause failure?

To obtain a confidence interval in general, we add and subtract a multiple of the appropriate standard deviation (for the mean that's the sem) to the statistic (e.g. $\bar{x}$) to obtain a range of values for the population parameter (e.g. μ) about which we have a pre-determined level of confidence.

We talk about this in terms of "confidence" rather than "probability", because we cannot attach a probability to the true parameter being contained in the specific range of values based on a single sample - the true parameter will either be in the interval or it won't.

B) CI for μ when σ^2 is known:

If $\overline{X}$ is our point estimate of μ then we can use $\sigma/\sqrt{n}$ to construct an interval and say "I am *w*% confident that μ is in this interval". An interval estimate may be preferable to a point estimate since it contains information about how variable the point estimate is.

We've learned that with large enough samples: $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, regardless of the underlying distribution of the X_i.

If we were to repeatedly sample from the population and form a CI each time, a desirable outcome would be to include the true parameter value μ in the collection of CI a high percentage of the time. We would like to say that the probability that μ is between the limits of the CI is *1- α.* The level of confidence we have is (1-α) x 100%. We call α the *level of significance*.

Our goal is to find a, b such that $P(a < \mu < b) = 1-\alpha$, where a and b are equidistant from μ.

For a standard normal distribution, here's how we begin:

$P(-Z_{1-\alpha/2} \leq Z \leq Z_{1-\alpha/2}) = 1-\alpha$.

What does this mean?

How do obtain a and b for any normal distribution?

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

$$\underbrace{P\left(-Z_{1-\alpha/2}\,\sigma/\sqrt{n} \le \bar{X}-\mu \le Z_{1-\alpha/2}\,\sigma/\sqrt{n}\right) = 1-\alpha}_{\text{multiply by } \sigma/n}$$

$$\underbrace{P\left(-\bar{X}-Z_{1-\alpha/2}\,\sigma/\sqrt{n} \le -\mu \le -\bar{X}+Z_{1-\alpha/2}\,\sigma/\sqrt{n}\right) = 1-\alpha}_{\text{subtract } \bar{X}}$$

$$\underbrace{P\left(\bar{X}+Z_{1-\alpha/2}\,\sigma/\sqrt{n} \ge \mu \ge \bar{X}-Z_{1-\alpha/2}\,\sigma/\sqrt{n}\right) = 1-\alpha}_{\text{multiply } -1 \text{ reverse} \le}$$

$$\underbrace{P\left(\bar{X}-Z_{1-\alpha/2}\,\sigma/\sqrt{n} \le \mu \le \bar{X}+Z_{1-\alpha/2}\,\sigma/\sqrt{n}\right) = 1-\alpha}_{\text{reverse}}$$

$$\text{for } \alpha = .05,\ Z_{1-\alpha/2} = 1.96$$

$$P\left(\bar{X}-1.96\,\sigma/\sqrt{n} \le \mu \le \bar{X}+1.96\,\sigma/\sqrt{n}\right) = .95$$

On repeated sampling, with $\alpha = 0.05$, μ will be contained between $\bar{X} \pm 1.96\,\sigma/\sqrt{n}$ (1-0.05) x 100%, or 95% of the time.

For a single, *specific* sample and value of $\bar{X}$, we say that we are 95% *confident* that the true mean is between $\bar{X} \pm 1.96\,\sigma/\sqrt{n}$, a 95% confidence interval for μ. The true mean will be in the specific interval or it won't, i.e. the probability that μ is in the interval is 0 or 1.

In general, $\bar{X} \pm Z_{1-\alpha/2}\,\sigma/\sqrt{n}$ is a (1-α) x 100% CI for μ.

Example - Figure 6.7 in Rosner.

Example. The 95% CI for μ the mean alcohol level at failure (assuming σ = 0.3 mg/ml) is $9.21 \pm 1.96(0.3/\sqrt{7})$ = (8.95 mg/ml, 9.43 mg/ml).

Summary: "We are 95% confident that mean alcohol level at failure is between 8.95 mg/ml and 9.42 mg/ml."

Other correct statements:

"At the 95% level of confidence, the data are consistent with a mean alcohol level at failure between 8.95 mg/ml and 9.42 mg/ml."

"On repeated sampling, 95% of the CI formed like this CI will contain the true value of mean alcohol level at failure." (Boring, uninformative, but correct …)

Let's examine the factors that influence the width of the CI for the population mean, μ: $\overline{X} \pm Z_{1-\alpha/2} \sigma/\sqrt{n}$

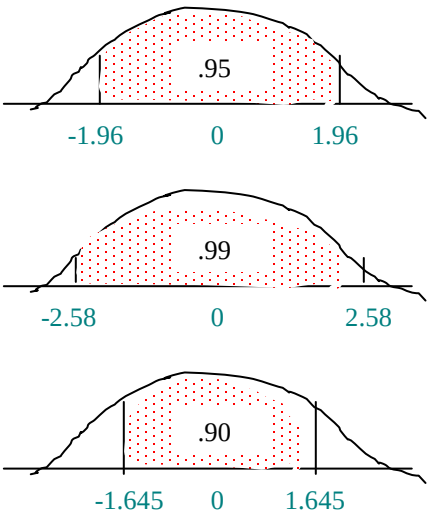
α

σ

n

For different levels of confidence, we can use SAS or R functions, the tables in Rosner, or Jack's Tables to replace 1.96 with the appropriate value:

95% CI	1.96
99% CI	2.58
90% CI	1.645



e.g. n = 22 normal adult females participating in a study of osteogenesis imperfecta had a mean oral temperature of $\bar{X}$ = 36.6° C with a standard deviation σ = 0.4°C and hence a sem: $0.4/\sqrt{22}$ = 0.09 °C

90% CI for μ: 36.6°C ± 1.645(0.09) = (36.452°C , 36.748°C)

Confidence statement:
"We are 90% confident that mean oral body temperature is between 36.45°C and 36.75°C."

Other correct statements:

Note: 95% confidence is the standard in most fields. Other levels of confidence can be used, depending on the context. This might then make it difficult to compare with other published results.

C) CI for μ when σ^2 unknown: the t-Distribution

In practice σ will *not* be known. So, when we standardize $\bar{X}$ we must use *s* in place of σ. This introduces more variation into $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ than a normal distribution has, so the standard

normal percentiles are no longer appropriate. Instead the quantity: $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$

has a *t-distribution* with n-1 degrees of freedom and we use *t*-tables, SAS or R functions, or Jack's Tables to replace the percentiles from the Z distribution (e.g. 1.96, 1.645). The percentiles of the *t*-distribution are slightly larger than the corresponding Z percentiles since the *t*-distribution has heavier tails than the normal distribution.

Let's look at Table 5 in the Rosner text, SAS Lab 4 in the section on probability functions and their inverses, and Jack's Tables.

For approximately normal sample means with unknown variance, a (1-α) x 100% CI for μ is: $\bar{X} \pm t_{n-1,1-\alpha/2} \, s/\sqrt{n}$.

Using the previous data, assuming σ is *not* known, a 90% CI for mean oral body temperature data is given by:

$\bar{X} \pm t_{22-1,\ 0.95} \, s/\sqrt{n} = 36.6 \pm 1.721(0.09) =$ (36.445°C , 36.755°C)

Note: This interval is a bit wider, reflecting our uncertainty about σ.

Confidence statement:

t-distribution (Student's *t*)

William Gossett, 1908 published "Student's *t*"
- Symmetric, bell-shaped, centered at 0
- More spread than normal; heavier tails
- There's a different *t*-distribution for each value of degrees of freedom (df)

Degrees of freedom: the (central) *t* distribution depends on one parameter called the degrees of freedom (df), which is equal to *n*-1. As with the chi-square distribution, the df are the number of independent data points. In this case it's the number of independent data points remaining after we estimate the mean μ using $\overline{X}$.

Assumptions for CI for the population mean, μ: There are four possibilities:

1. *n* small; σ known; data $X_1, \ldots, X_n$ - normal dis'n
X_i data normal, so $\bar{X}$ exactly normal; σ known so use Z tables; this is a rare scenario

2. *n* small; σ unknown; data $X_1, \ldots, X_n$ – normal dis'n
X_i data normal, so $\bar{X}$ exactly normal; σ unknown so use t to account for estimating σ by s; this is a common scenario

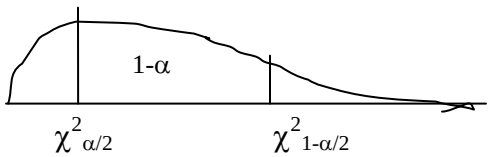
3. *n* small; data $X_1, \ldots, X_n$ - not from normal dis'n
X_i data not normal, for *n* small we can't use CLT to say $\bar{X}$ is approximately normal; hard to justify using Z- or *t*-based CI; alternative: nonparametric methods – later this semester; you will see the t-disn used in this scenario

4. *n* large-ish (≥30); data $X_1, \ldots, X_n$ not normal

Even if X_i data are not normal, CLT says $\bar{X}$ ***is*** approximately normal; if σ is unknown, use *t*. If σ is known (almost never will be) can use Z. We'll see very little difference since $t_{df,\ 1-\alpha/2} \rightarrow Z_{1-\alpha/2}$ as $df \rightarrow \infty$

D) CI for the population variance, σ^2

Recall: If $X_1, X_2, \ldots, X_n$ iid N [μ , σ^2] then $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$



$$1-\alpha = P\left(\chi^2_{n-1,\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{n-1,1-\alpha/2} \right)$$

$$= P\left(\frac{\chi^2_{n-1,\alpha/2}}{(n-1)s^2} \le \frac{1}{\sigma^2} \le \frac{\chi^2_{n-1,1-\alpha/2}}{(n-1)s^2} \right)$$

$$= \underbrace{P\left(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \ge \sigma^2 \ge \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \right)}_{reciprocal}$$

$$= \underbrace{P\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \right)}_{rearrange}$$

$\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}$ and $\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}$ are the (1-α) x 100% confidence limits for σ^2

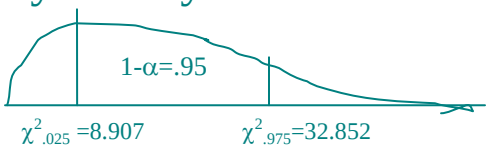
$\sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}}$ and $\sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}}$ are the (1-α) x 100% confidence limits for

σ

e.g. X = age in general population; n = 20 $s^2 = 9$ years^2

What is the 95% CI for σ^2? $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) = \chi^2(19)$

Note asymmetry:



Lower Limit: $\frac{(n-1)s^2}{\chi^2_{19,.975}} = \frac{(19)9}{32.852} = 5.205 \text{ years}^2$

Upper Limit: $\frac{(n-1)s^2}{\chi^2_{19,.025}} = \frac{(19)9}{8.907} = 19.198 \text{ years}^2$

Confidence statement:
We are 95% confident that the true pop'n variance is between 5.2 yrs^2 and 19.2 yrs^2.

At a 95% confidence level the data is consistent with true pop'n variance lying between 5.2 yrs^2 and 19.2yrs^2.

E) CI for Binomial Proportions – we use $\hat{p}$ to estimate p.

There are two ways of obtaining a CI for p:

1) Apply the normal approximation to the binomial

We know that $\mu_{\hat{p}} = p$ and $\sigma^2_{\hat{p}} = \frac{pq}{n}$.

If $n\hat{p}\hat{q} \geq 5$, the normal approximation to the binomial should apply and we can standardize to obtain $\frac{\hat{p}-p}{\sqrt{\hat{p}\hat{q}/n}} \sim N(0,1)$.

Then, using the same principles for obtaining a CI for the sample mean, a (1-α) x 100% confidence interval for *p* is given by:

$$\hat{p} \pm Z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n}\ .$$

e.g. n = 25, X = number alive after 1 year of treatment for colon cancer = 8, $\hat{p} = 8/25 = 0.32$. First, check to see if $n\hat{p}\hat{q} \geq 5 : 25(0.32)(0.68) = 5.44$; thus, the normal approximation holds.

To obtain a 95% CI for p:

$$\hat{p} \pm Z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n} = 0.32 \pm 1.96\sqrt{(0.32)(0.68)/25} = \quad (0.137, 0.503)$$

Confidence statement: With confidence level of 95%, the true proportion of patients surviving after 1 year of treatment is in the interval (0.137, 0.503).

2) Exact Method (for small or large *n*)

We can use exact binomial probabilities to obtain a confidence interval for *p:* (p_1, p_2). The lower limit p_1 of the CI must satisfy the following:

$$p_1 : P(X \geq x | p = p_1) = \frac{\alpha}{2} = \sum_{k=x}^{n} \binom{n}{k} p_1^k (1-p_1)^{n-k}$$

This is the probability of observing outcomes ≥ the number observed in the data. The upper limit p_2 of the CI must satisfy the following:

$$p_2 : P(X \leq x | p = p_2) = \frac{\alpha}{2} = \sum_{k=0}^{x} \binom{n}{k} p_2^k (1-p_2)^{n-k},$$

the probability of observing outcomes ≤ the number observed in the data. Note: The exact limits are not usually symmetric around the point estimate, but they are symmetric with the normal approximation-based CI.

These limits can be difficult to compute. For a limited number of sample sizes, there are special tables (nomograms) that can be used for this purpose – see Table 7a and 7b in the Rosner text. Using them however requires eyeballing the values, not such an exact method after all!

An alternative is to use SAS PROC FREQ. With appropriate code you can find the limits for any *x* and *n*. Also, through a series of programming statements involving the PROBBNML function in SAS you can find the limits for any *x* and *n*. This latter approach better illustrates what the formulas on p. 11 mean.

e.g. n = 20, X = number alive after 1 year treatment for colon cancer = 8, $\hat{p} = 8/20 = 0.4$, $n\hat{p}\hat{q} = 20(0.4)(0.6) = 4.8 < 5$

Using SAS PROC FREQ:

```
DATA one;
INPUT alive wt;
CARDS;
1 8
0 12
;
RUN;

PROC FORMAT;
VALUE survive 0 = 'No' 1 = 'Yes';
RUN;

PROC FREQ DATA=one;
TABLES alive / BINOMIAL (LEVEL = 'Yes') ALPHA = .05;
WEIGHT wt;
FORMAT alive survive.;
RUN;
```

The SAS System

The FREQ Procedure

alive	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	12	60.00	12	60.00
Yes	8	40.00	20	100.00

Binomial Proportion for alive = Yes	
Proportion	0.4000
ASE	0.1095
95% Lower Conf Limit	0.1853
95% Upper Conf Limit	0.6147

Exact Conf Limits	
95% Lower Conf Limit	0.1912
95% Upper Conf Limit	0.6395

Test of H0: Proportion = 0.5	
ASE under H0	0.1118
Z	-0.8944
One-sided Pr < Z	0.1855
Two-sided Pr > |Z|	0.3711

Sample Size = 20

This page: my homegrown program in SAS to find exact binomial CI - not pretty but it works. ☺

```
data one;
  n = 20; * Given sample size;
  do x = 0 to n; * Enumerate possible outcomes for r.v. x;
    phat = x/n; * Estimate of phat for each outcome;
    do p = 0.001 to .999 by .001; * Enumerate possible values of true
proportion p;
      if x = 0 then aprob=1;
      else aprob = 1-probbnml(p, n, x-1); * Upper tail binomial probability
based on p, n, x;

      if x = n then bprob = 0;
      else bprob = probbnml(p, n, x); * Lower tail binomial probability
based on p, n, x;

    output;
    end;
  end;
run;

data two; set one;

  if aprob le .025; * Select scenarios with the correct upper tail binomial
probability;
  conf = .95;

  p1 = p;
  run;

proc sort data=two; by phat p1;

data three; set two; by phat p1;
  if last.phat;
run;

data four; set one;
  if bprob le .025; * Select scenarios with the correct lower tail binomial
probability;
  conf = .95;

  p2 = p;
run;

proc sort data=four; by phat p2; run;

data five; set four; by phat p2;
  if first.phat;
run;

data all;
  merge three five; by phat; * This is the set of scenarios with the
correct upper AND lower tail binomial probabilities;
  if p1 = . then p1 = 0;
  if p2 = 0.001 or p2 = . then p2 = 1.0;
  run;

proc print data=all;
 where phat = 0.4; * This will give the values for the class example;
 var n phat p1 p2 conf;
run;
```

Obs	n	phat	p1	p2	conf
9	20	0.40	0.191	0.640	0.95

(Compare these limits to the Exact ones given above by SAS.)

F) One-sided CI

For some questions that we will want to address, one-sided confidence intervals will be desirable instead of the two-sided ones we've discussed so far.

Upper one-sided CI:

(a, ∞): a is the lower bound

e.g. quality control: we want to be assured that *at least* a certain threshold is achieved (a minimum quantity exists)

Lower one-sided CI:

(-∞, b): b is the upper bound

e.g. ensure that a toxic substance *does not exceed* a maximum, b

e.g. ensure variability in data is *no larger than* b

Type of CI	Upper CI (lower bound)	Lower CI (upper bound)
μ with σ^2 known	$\bar{X} - Z_{1-\alpha} \sigma/\sqrt{n} < \mu$	$\mu < \bar{X} + Z_{1-\alpha} \sigma/\sqrt{n}$
μ with σ^2 unknown	$\bar{X} - t_{n-1,1-\alpha} s/\sqrt{n} < \mu$	$\mu < \bar{X} + t_{n-1,1-\alpha} s/\sqrt{n}$
σ^2	$\frac{(n-1)s^2}{\chi^2_{1-\alpha}} < \sigma^2$	$\sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha}}$
p	$\hat{p} - Z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n} < p$ When the normal approximation doesn't apply -> SAS or R	$p < \hat{p} + Z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}$ When the normal approximation doesn't apply -> SAS or R

Example – A standard therapy exists for a certain type of cancer and patients receiving that regimen show a 5-year survival rate of 30%. A new therapy is being tested and it has an unknown 5-year survival rate. We would only be interested in pursuing use of the new therapy if it were better than the standard.

Suppose 40 out of 100 patients who receive the new therapy survive for 5 years. Can we say that the new therapy is better than the standard?

For 95% one-sided CI: Z percentile = 1.645
For 97.5% one-sided CI: Z percentile = 1.96

Upper one-sided 95% CI = 0.4 – 1.645*sqrt(0.4*0.6/100) = (0.319, 1)

Upper one-sided 97.5% CI = 0.4 – 1.96*sqrt(0.4*0.6/100) = (0.303, 1)

Conclusion:

Prediction Intervals: - prediction for future observations

Example. Birthweight (g) from a sample of 10 babies born at Boston City Hospital: 116, 124, 119, 100, 127, 103, 140, 82, 107, 132

a. Give an interval that is likely to contain the mean birthweight for the population – *a confidence interval*

b. Give an interval that is likely to contain the weight of the next baby born from the population – *a prediction interval*

(1-α) x 100% CI for μ: $\bar{X} \pm t_{n-1,1-\alpha/2}\left(s\sqrt{1/n}\right)$

e.g. We're 95% confident that this interval contains the true mean μ.

(1-α) x 100% PI for μ: $\bar{X} \pm t_{n-1,1-\alpha/2}\left(s\sqrt{\underbrace{1}_{\text{spread in individual data values}} + \underbrace{1/n}_{\text{estimating } \mu}}\right)$

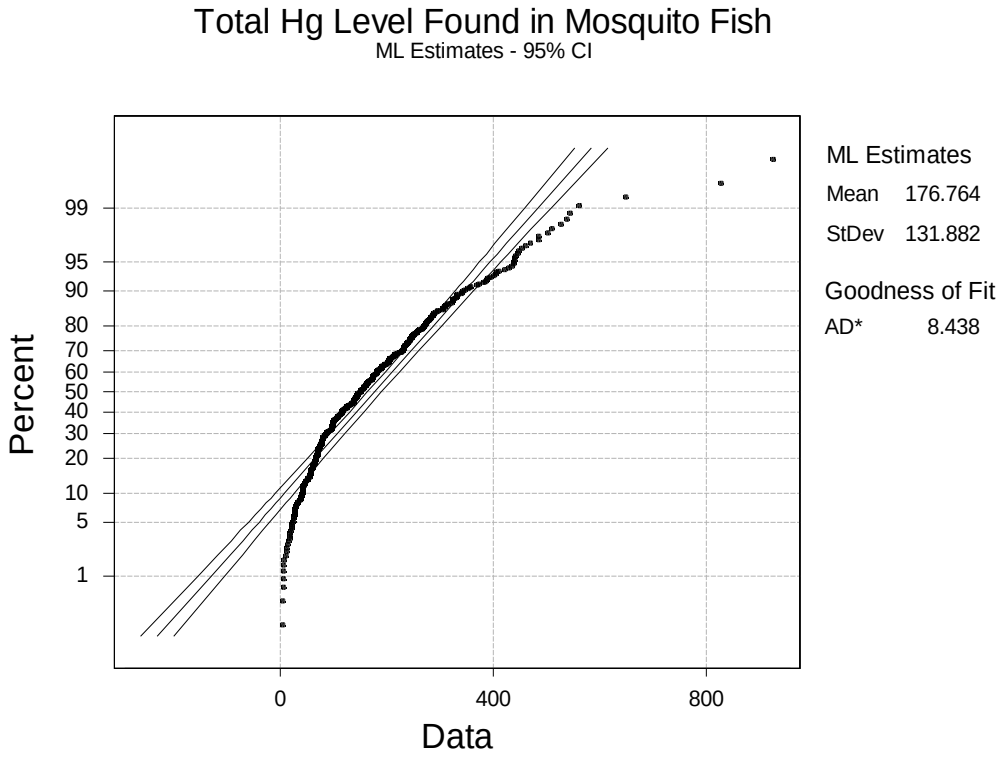
e.g. We're 95% confident that this interval contains the weight of the next individual baby to be born.

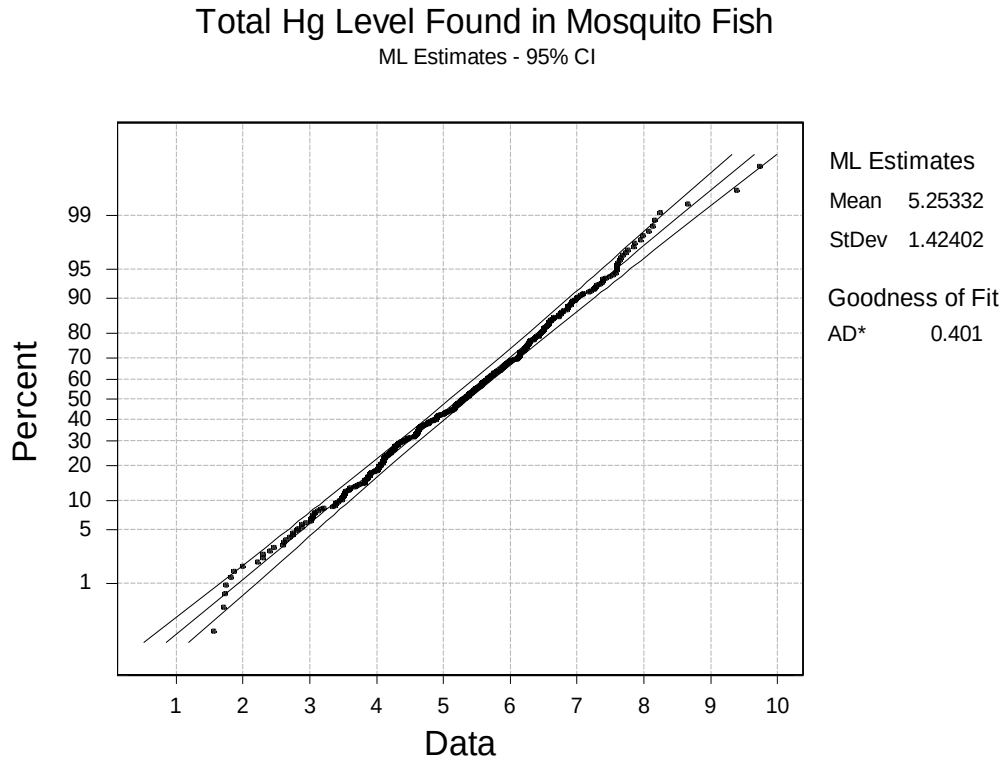
For birthweight sample, $\bar{X}$ = 115oz, s = 17.25oz, $t_{9,0.975}$ = 2.262

95% CI for μ, the mean birthweight of the population: 115oz ± 12.3 or (102.7oz, 127.5oz)

95% PI for the birthweight of a single infant: 115oz ± 40.9 or (74.1oz, 155.9oz)

Confidence intervals for normal probability plots – from Desireé Bailey, MS (received her degree from UCD/AMC in biostatistics).





The difference between these is a $\log_2$ transformation.

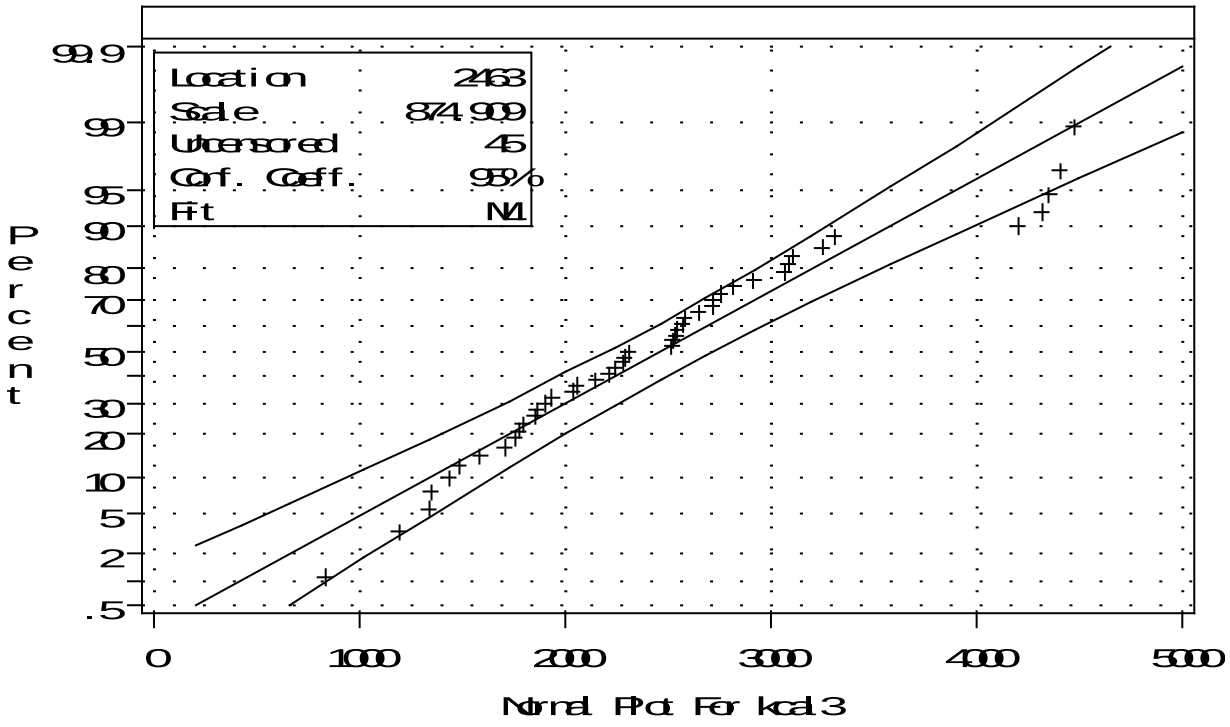
Let's apply this idea and thinking to the diet data from Lectures 1 and 9:

```
data diet;
  set diet;
  status = 0;
run;

proc reliability data=diet;
  distribution Normal;
  pplot kcal3*Status(1); /* no censored values */
run;
```

Note: The axes are reversed from the way the normal probability plot is created by SAS PROC UNIVARIATE.

Do the data follow a normal distribution?



William Sealey Gosset

William Sealy Gosset (June 13, 1876 – October 16, 1937) was a chemist and statistician, better known by his pen name ***Student***. Born Canterbury, England to Agnes Sealy Vidal and Colonel Frederic Gosset, Gosset attended Winchester College, the famous private school, before reading chemistry and mathematics at New College, Oxford. On graduating in 1899, he joined the Dublin brewery of Arthur Guinness & Son.

Guinness was a progressive agro-chemical business and Gosset would apply his statistical knowledge both in the brewery and on the farm—to the selection of the best yielding varieties of barley. Gosset acquired that knowledge by study, trial and error and by spending two terms in 1906/7 in the biometric laboratory of Karl Pearson. Gosset and Pearson had a good relationship and Pearson helped Gosset with the mathematics of his papers. Pearson helped with the 1908 papers but he had little appreciation of their importance. The papers addressed the brewer's concern with small samples, while the biometrician typically had hundreds of observations and saw no urgency in developing small-sample methods.

Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness brewery. To prevent further disclosure of confidential information, Guinness prohibited its employees from publishing any papers regardless of the contained information. This means that Gosset was unable to publish his works under his own name. Therefore he used the pseudonym *Student* for his publications to avoid detection of his publications by his employer. Therefore his most famous achievement is now referred to as the Student t-distribution, which may otherwise have been the Gosset t-distribution.

Using this pseudonym Pearson published *The probable error of a mean* and almost all of Gosset's papers in his journal *Biometrika*. However, it was Ronald Fisher who appreciated the importance of Gosset's small-sample work, after Gosset had written to him to say *I am sending you a copy of Student's Tables as you are the only man that's ever likely to use them!*. Fisher believed that Gosset had effected a "logical revolution". Ironically the *t*-statistic for which Gosset is famous was actually Fisher's creation. Gosset's statistic was $z = t/\sqrt{(n - 1)}$. Fisher introduced the *t*-form because it fitted in with his theory of degrees of freedom. Fisher was also responsible for the applications of the *t*-distribution to regression.

Although introduced by others, Studentized residuals are named in Student's honor because, like the problem that led to Student's t-distribution, the idea of adjusting for estimated standard deviations is central to that concept.

Gosset's interest in barley cultivation led him to speculate that design of experiments should aim, not only at improving the average yield, but also at breeding varieties whose yield was insensitive (robust) to variation in soil and climate. This principle only occurs in the later thought of Fisher and then in the work of Genichi Taguchi in the 1950s.

In 1935, he left Dublin to take up the position of Head Brewer, in charge of the scientific side of production, at a new Guinness brewery in London. He died in Beaconsfield, England.

Gosset was a friend of both Pearson and Fisher, an achievement, for each had a massive ego and a loathing for the other. Gosset was a modest man who cut short an admirer with the comment that "Fisher would have discovered it all anyway."

Posted on the course website under Files -> Papers : **Student (1907). On the error of counting with a haemacytometer. *Biometrika,* 5:351-360. Student (1908). The probable error of a mean. *Biometrika,* 6:1-25.**

B. CI for μ when σ^2 known.

```
# The z value for the standard normal central 95%
z95 <- c(qnorm(0.025), qnorm(0.975))
z95

[1] -1.96  1.96

ci.95 <- 9.21 + z95 * (0.3/sqrt(7))
ci.95

[1] 8.9878 9.4322

z90 <- c(qnorm(0.05), qnorm(0.95))
z90

[1] -1.6449  1.6449

ci.90 <- 36.6 + z90 * 0.09
ci.90

[1] 36.452 36.748
```

C. CI for μ when σ^2 unknown: t-distribution.

```
t90 <- c(qt(0.05, df = 21), qt(0.95, df = 21))
t90

[1] -1.7207  1.7207

ci <- 36.6 + t90 * (0.09)
ci

[1] 36.445 36.755
```

D. CI for the population variance σ^2: χ^2 distribution.

```
chi95 <- c(qchisq(0.025, df = 19), qchisq(0.975, df = 19))
chi95

[1]  8.9065 32.8523

ci.chi <- 19 * 9/chi95
ci.chi

[1] 19.1994  5.2051
```

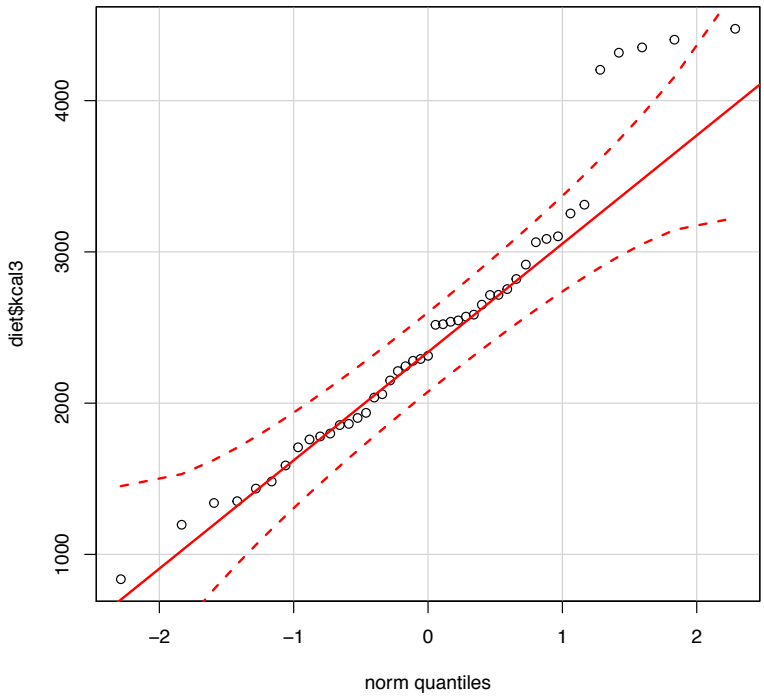
E. Exact method for binomial confidence interval using package `Hmisc`.

```
library(Hmisc)
binconf(8, 20, method = "exact")

 PointEst    Lower    Upper
      0.4 0.19119 0.63946
```

Cofidence interval for Normal probability plot using package `car`.

```
library(car)
diet <- read.csv("~/Dropbox/6611METHODS/6611/diet.csv")
qqPlot(diet$kcal3)
```



Appendix: Code

```
# unit 11
## ---- ex1 ----

# The z value for the standard normal
# central 95%
z95<-c(qnorm(0.025),qnorm(0.975))
z95
ci.95<-9.21+z95*(.3/sqrt(7));ci.95

z90<-c(qnorm(0.05),qnorm(0.95))
z90
ci.90<-36.6+z90*0.09;ci.90

## ---- ex2 ----
t90<-c(qt(0.05,df=21),qt(0.95,df=21))
t90
ci<-36.6+t90*(0.09);ci

## ---- ex3 ----
chi95<-c(qchisq(0.025,df=19),qchisq(0.975,df=19))
chi95
ci.chi<-19*9/chi95
ci.chi

## ---- ex4 ----
library(Hmisc)
binconf(8,20,method="exact")

## ---- ex5 ----
library(car)
diet <- read.csv("~/Dropbox/6611METHODS/6611/diet.csv")
qqPlot(diet$kcal3)
```