Semester 2, 2012 (Last adjustments: September 10, 2012)

Lecture Notes

MATH1905 Statistics (Advanced)

Lecturer

Dr. John T. Ormerod

School of Mathematics & Statistics F07

University of Sydney

(w) 02 9351 5883

(e) john.ormerod (at) sydney.edu.au

Semester 1, 2012 (Last adjustments: September 10, 2012)

Monday, 10th September 2012

Lecture 1 - Content

- ☐ **Joint distributions**
- ☐ **Independent random variables**
- ☐ **Central limit theorem**

References from Phipps & Quine

- ☐ Section 2.4 pages 69-72.
- ☐ Section 3.2 pages 73-75.

Standard Normal Distribution

Let $Z \sim N(0,1)$ then

- The probability density function at z is given by

  ```
  > dnorm(z)
  ```

- The (cumulative) distribution function at z, $\Phi(z) = \mathrm{P}(Z < z)$, is given by

  ```
  > pnorm(z)
  ```

- The inverse (cumulative) distribution function at t, $\Phi^{-1}(t)$ or the value of z such that $\Phi(z) = t$, is given by

  ```
  > qnorm(t)
  ```

- To generate n random values from $Z \sim N(0,1)$ we use

  ```
  > rnorm(n)
  ```


Normal Distribution

Let $X \sim N(\mu, \sigma^2)$ then

- The probability density function at x is given by

  ```
  > dnorm(x,mu,sigma)
  ```

- The (cumulative) distribution function at x, $\Phi((x-\mu)/\sigma) = \mathrm{P}(X < x)$, is given by

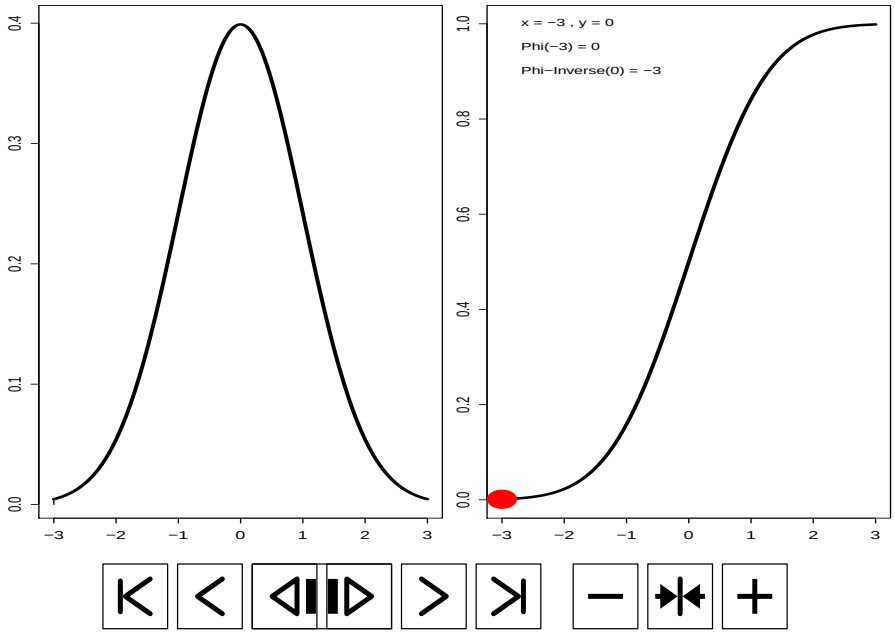
  ```
  > pnorm(x,mu,sigma) # OR
  > pnorm( (x-mu)/sigma )
  ```

- The inverse (cumulative) distribution function at t, the value of x such that $\mathrm{P}(X < x) = t$, is given by

  ```
  > qnorm(t,mu,sigma)
  ```

- To generate n random values from $X \sim N(\mu, \sigma^2)$ we use

  ```
  > rnorm(n,mu,sigma)
  ```


x = −3 , y = 0
Phi(−3) = 0
Phi−Inverse(0) = −3

Joint distributions

Independence of random variables

Let X be a real-valued random variable (e.g. normal, exponential, binomial) and $x \in \mathbb{R}$ any number, then

$$A = \{X \leq x\}$$

represents an event. Let Y be another real-valued random variable and

$$B = \{Y \leq y\}, \quad y \in \mathbb{R}.$$

Recall the definition of independence of events: A and B are independent iff

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$$

which is a special case of the general multiplication rule,

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B|A) = \mathrm{P}(B)\,\mathrm{P}(A|B) \quad \text{if} \quad \mathrm{P}(A), \mathrm{P}(B) \neq 0.$$

Definition 1. Two random variables X and Y are independent if and only if for any numbers x and y the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent events.

Example.

- ($X =$ 'height', $Y =$ 'weight') from a random person are not independent.
- $X_1 =$ 'lottery numbers next draw' and $X_2 =$ 'lottery numbers in three weeks time' are
- $X_1 =$ 'todays rainfall' and $X_2 =$ 'tomorrows rainfall' are

From the above Definition 1 we easily get the joint cumulative distribution function and joint probability density function of independent random variables.

Joint distribution functions and densities

Definition 2. The joint cumulative distribution function of two random variables X and Y is

$$F_{X,Y}(x, y) := \mathrm{P}(X \leq x, Y \leq y)$$

and the joint density function is denoted $f_{X,Y}(x, y)$.

Note that, if X and Y are continuous random variables, then $F_{X,Y}(x, y)$ and $f_{X,Y}(x, y)$ are related via

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}$$

and

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s, t) dt ds.$$

Marginal distribution functions and densities

Definition 3. If $F_{X,Y}(x,y)$ is the joint cumulative distribution function of two random variables X and Y then, $F_X(x)$ and $F_Y(y)$ are called the marginal cumulative distribution functions of X and Y, respectively.

For integer valued random variables the marginal probability mass functions can be calculated via

$$P(X=x)=\sum_y P(X=x,Y=y) \qquad \text{and} \qquad P(Y=y)=\sum_x P(X=x,Y=y)$$

while for continuous random variables the marginal density functions can be calculated via

$$f_X(x)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \qquad \text{and} \qquad f_Y(y)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx.$$

From these the marginal cumulative distribution functions can be calculated in the usual way.

Expectations of Joint Distributions

Let $g(x,y)$ be a bivariate function and let X and Y be random variables with joint density function $f_{X,Y}(x,y)$.

If X and Y are discrete random variables then

$$E[g(X,Y)]=\sum_x\sum_y g(x,y)f_{X,Y}(x,y).$$

If X and Y are continuous random variables then

$$E[g(X,Y)]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dxdy.$$

Independence

Definition 4. Let $F_X(x) = \mathrm{P}(X \leq x)$ and $F_Y(y) = \mathrm{P}(Y \leq y)$ be the cumulative distribution functions of the independent random variables X and Y then, the joint cumulative distribution function is

$$F_{X,Y}(x, y) := \mathrm{P}(X \leq x, Y \leq y) = \mathrm{P}(X \leq x)\,\mathrm{P}(Y \leq y) = F_X(x)F_Y(y).$$

Definition 5. Let $f_X(x)$ and $f_Y(y)$ be the probability density functions of the independent random variables X and Y then, the joint probability density function is given by

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

Independent random variables: rules for expectations and variances

Theorem 1 (Properties of E and Var)**.** Let X and Y be random variables then

1. $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$
2. if X and Y are independent then, $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$
3. if X and Y are independent then, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Note that for any two, not necessarily independent, random variables

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)$$

where

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))].$$

Proof of 1. discrete case only:

Central limit theorem

Many observed phenomena can be modelled as the sum of several random variables:

- total weight of passengers in a lift,
- total of available funds

or means of random variables

- average class mark,
- average height and weight,
- average temperature in Sydney.

The central limit theorem is useful in these types of situations.

Some useful facts about the normal distribution

Theorem 2. Let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, let X and Y be independent and let a and b be two real numbers. Then

$$Z = aX + bY \sim \mathcal{N}(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2).$$

Proof: Not in MATH 1905.
In general, let $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ be independent and a_i be real numbers for $1 \leq i \leq n$ then

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

Example

Example (Mean and variance of the sample mean $\overline{X}$)**.** Let the n random variables $X_1, X_2, \ldots, X_n$ be pairwise independent and each have the same distribution with mean μ and variance σ^2. Then for the sample mean, that is $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we have,

i) mean: $\mu_{\overline{X}} = E(\overline{X}) = \mu$

ii) variance: $\sigma_{\overline{X}}^2 = \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$

Sums of normal random variables

Theorem 3. If all $X_i \sim \mathcal{N}(\mu, \sigma^2)$ then,

$$T = \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{and} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

[This is only true for normal rvs; in STAT2911 moment generating functions are introduced that make a simple proof available].

Example. X_1, X_2, X_3 are independent random variables with

i	0	1	3
p_i	$1/3$	$1/3$	$1/3$

$\overset{T_2=X_1+X_2}{\Rightarrow}$

i	0	1	2	3	4	6
p_i	$1/9$	$2/9$	$1/9$	$2/9$	$2/9$	$1/9$

$\Rightarrow$

i	0	1	2	3	4	5	6	7	9
p_i	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{4}{27}$	$\frac{6}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{1}{27}$

(Note, the distribution of T_3 clusters around the mean $\mathrm{E}\,T_3 = 4$.)

Theorem 4 (CLT, central limit theorem)**.** If $X_1, X_2, \ldots, X_n$ are iid random variables with mean μ and variance $0 < \sigma^2 < \infty$ then,

$$\mathrm{P}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \leq x\right) \to \Phi(x) = \mathrm{P}(Z \leq x) \quad \text{as} \quad n \to \infty.$$

Proof. Postponed to second year... □

Thus for n large (here $n \geq 25$) the following are approximately true:

$$T = \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$
$$\overline{X} = \frac{1}{n}T \sim \mathcal{N}(\mu, \sigma^2/n).$$

The closer the distribution of X_i is to the normal the better the approximation for small n values.

Example (PQ, p71)**.** Steel rods, made with diameter $R \sim \mathcal{N}(4.90, 0.03^2)$ (in cm), are to fit into sockets, made with diameter $S \sim \mathcal{N}(5.00, 0.04^2)$ (in cm). For a satisfactory fit the socket diameter should exceed the rod diameter, but by no more than 0.20 cm. If a rod and socket are taken at random, what is the probability that the fit is unsatisfactory?

Example. The tibia length of a certain species of beetle can be modelled by $L \sim \mathcal{N}(7.8, 0.3^2)$ mm.

(i) What is the probability that the average length of 25 independent tibia lengths will be less than 7.6 mm?

Solution (i):

Because of the CLT the answer will be approximately correct regardless of the exact distribution of tibia length.

Example. The tibia length of a certain species of beetle can be modelled by $L \sim \mathcal{N}(7.8, 0.3^2)$ mm.

(ii) What is the prob. that the average will differ from 7.8 by more than 0.1?

Solution (ii):

Note we can show that

$$\mathrm{P}(|L - 7.8| > 0.1) = 0.7414$$

so the average varies much less than the individual measurements.

Again, because of the CLT the answer will be approximately correct regardless of the exact distribution of tibia length.

Example. Systolic blood pressure readings for pre-menopausal, non-pregnant women aged 35 – 40 have $\mu = 122.6$ mm Hg and an s.d. of 11 mm Hg. An independent sample of 25 women is drawn from this target population and their BP recorded.

(i) What is the probability that the average BP is greater than 125 mm hg?

Solution (i):

Example. Systolic blood pressure readings for pre-menopausal, non-pregnant women aged 35 – 40 have $\mu = 122.6$ mm Hg and an s.d. of 11 mm Hg. An independent sample of 25 women is drawn from this target population and their BP recorded.

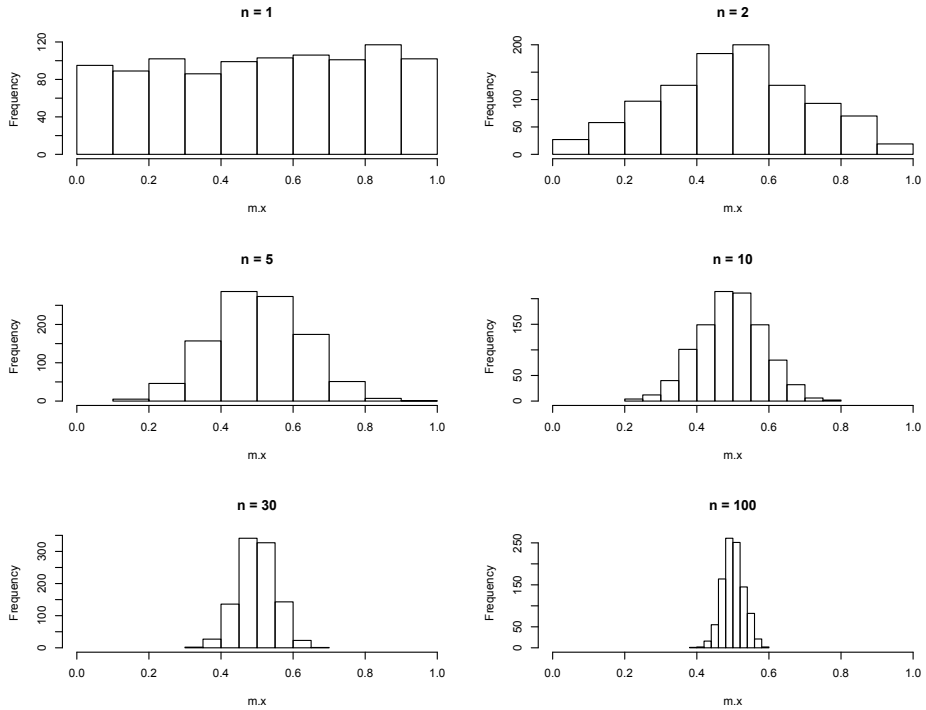
(ii) If the sample size increases to 40 how changes the answer to (i)?

Solution (ii):

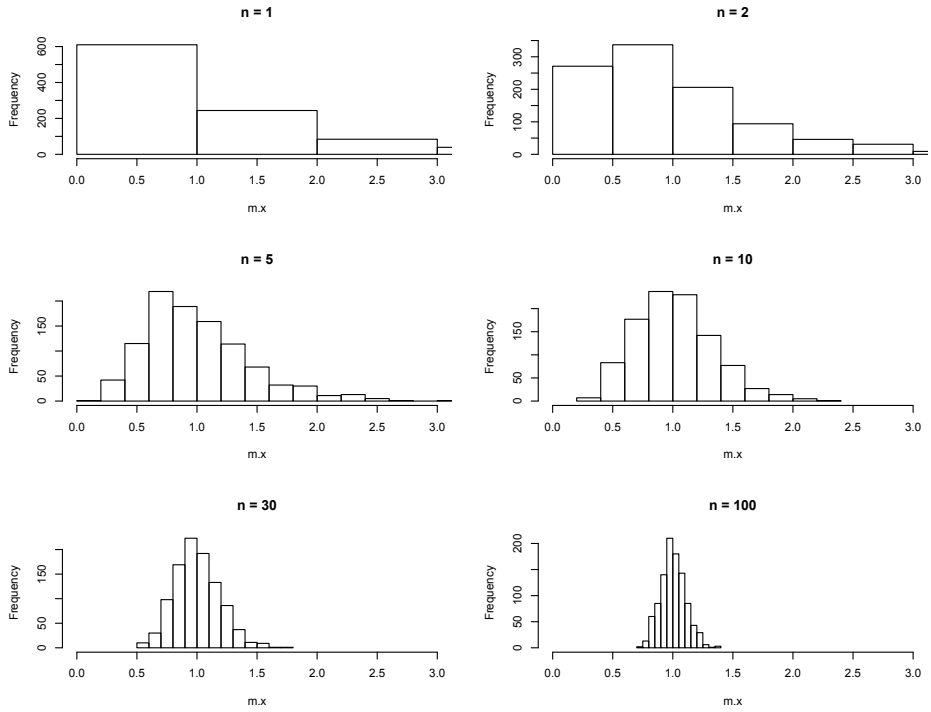
CLT in R

```
# Uniform distribution and CLT
n = 1; Loops = 1000; m.x = rep(0,Loops)
for(i in 1:1000){ x = runif(n); m.x[i] = mean(x)}
hist(m.x,xlim=c(0,1),main="n = 1")


# Exponential distribution and CLT
par(mfrow=c(3,2))
n = 1; Loops = 1000; m.x = rep(0,Loops)
for(i in 1:1000){ x = rexp(n); m.x[i] = mean(x)}
hist(m.x,xlim=c(0,3),main="n = 1")

# choose n in {1,2,5,10,30,100}

qqnorm(m.x)          # produces a Q-Q-plot
plot(density(m.x))   # produces a smooth estimated density
```


CLT for $X \sim \mathcal{U}$



CLT for $X \sim \mathcal{E}$



More on functions of random variables

In many applications interest focuses on some function $g(X)$ of the random variable X. E.g. change scale from meters to millimeters, logarithm of daily exchange rate changes or squared body height (BMI). In STAT2911 you will learn more. In the following I show some results that you are likely to understand with what you already know.

Theorem. (A simple version of the transformation theorem for densities) Let the random variable X have probability density f_X and let $g : \mathbb{R} \to \mathbb{R}$ be some monotone function and $h = g^{-1}$ be the inverse of g with

$$\frac{\partial h(y)}{\partial y} = h'(y).$$

Then, the density function of $Y = g(X)$ is given by $f_Y(y) = f_X(h(y)) \cdot |h'(y)| \cdot 1_{g(\mathbb{R})}(y)$.

Proof. From the definition of the probability density of Y and by applying the chain rule we get for a non-decreasing function g, that

$$\begin{aligned} f_Y(y) &= \frac{\partial}{\partial y} F_Y(y) = \frac{\partial}{\partial y} \mathrm{P}(Y \le y) \\ &= \frac{\partial}{\partial y} \mathrm{P}(g(X) \le y) = \frac{\partial}{\partial y} \mathrm{P}(X \le h(y)) \\ &= \frac{\partial}{\partial y} F_X(h(y)) = f_X(h(y)) h'(y) \\ &= f_X(h(y)) \cdot |h'|. \end{aligned}$$

For a non-increasing function $g(\cdot)$ the proof is essentially the same. □

Example. Let $X \sim \mathcal{U}(0,1)$ and $Y = g(X) = X^c, \quad c > 0$. The inverse of g equals $h(y) = y^{\frac{1}{c}}$, its derivative is $\frac{\partial h(y)}{\partial y} = \frac{1}{c} \cdot y^{\frac{1}{c}-1}$. From the transformation theorem it follows that

$$f_Y(y) = f_X(h(y)) \cdot |h'(y)| = 1 \cdot \frac{1}{c} \cdot y^{\frac{1}{c}-1} \cdot 1_{(0,1)}(y).$$

Lecture 2 - Content

- Normal approximation to the Binomial
- Sampling distributions

References from Phipps & Quine

- Section 3.1 pages 72-73.
- Section 3.3 pages 75-78.

Normal approximation to the Binomial

Let X_i be independent random variables (outcomes of Bernoulli trials), defined as

$$X_i = \begin{cases} 1 & \text{if the } i\text{th trial is a S,} \\ 0 & \text{if the } i\text{th trial is a F,} \end{cases}$$

and let $p = \mathrm{P}(\mathsf{S})$ on the ith trial.

Theorem 5. Let $X = X_1 + \ldots + X_n \sim \mathcal{B}(n, p)$ with $\mathrm{E}(X) = np$ and $\mathrm{Var}(X) = n\,\mathrm{Var}(X_1) = np(1-p)$. Then, X is approximately $\mathcal{N}(np, np(1-p))$.

Proof. Postponed to second year... □

- The approximation is quite good if $np \geq 5$ and $n(1-p) \geq 5$!
- The closer p is to 0.5 the better the approximation for small n.

Example. $(X \sim \mathcal{B}(12, 0.5))$

$$\mathrm{P}(X = 3) = \binom{12}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^9 = \frac{12 \times 11 \times 10}{1 \times 2 \times 3} \cdot \frac{1}{2^{12}} = 0.0537.$$

Comparing to the area under the approximating normal curve, e.g.

$$X \simeq Y \sim \mathcal{N}(6, 3)$$

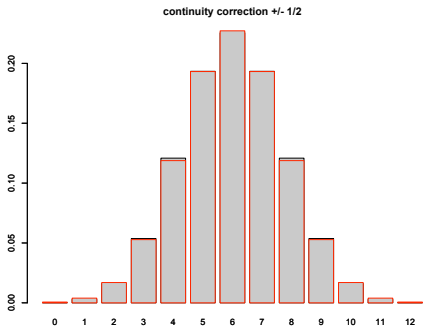
$$\mathrm{P}(X = 3) \simeq \mathrm{P}\left(3 - \lambda < Y < 3 + (1 - \lambda)\right); \quad \lambda \in [0, 1].$$

Most authors choose $\lambda = 1/2$ which in the above example is clearly closer to the true value of $\mathrm{P}(X = 3)$:

```
> mu=6; sd=sqrt(3);
> pnorm(4.0,mu,sd) - pnorm(3.0,mu,sd)
[1] 0.08247428
> pnorm(3.5,mu,sd) - pnorm(2.5,mu,sd)
[1] 0.05280327
> pnorm(3.0,mu,sd) - pnorm(2.0,mu,sd)
[1] 0.03117159
```


Example (cont)

- Note that `pnorm(3.5176,mu,sd) - pnorm(2.5176,mu,sd)` comes very close to `dbinom(3,12,0.5)` but is only optimal for this particular example.
- Overall performance of $\lambda = 1/2$ is best.



Continuity correction

- To approximate binomial probabilities using the normal consider areas of corresponding rectangles.
- Adjust the normal probability statement by adding or subtracting 0.5 to the constant to increase the area under the normal curve.

$$\begin{aligned} \mathrm{P}(X = x) &\simeq \mathrm{P}(x - 0.5 < Y < x + 0.5) \\ &= \mathrm{P}\left(\underbrace{\frac{x - 0.5 - \mu}{\sigma}}_{z_l} < Z < \underbrace{\frac{x + 0.5 - \mu}{\sigma}}_{z_u} \right) \\ &= \Phi(z_u) - \Phi(z_l). \end{aligned}$$

- For $\mathrm{P}(X \geq x)$ repeat the above step by noting:

$$\mathrm{P}(X \geq x) = \sum_{i \geq x} \mathrm{P}(X = i).$$

Example. If $X \sim \mathcal{B}(12, 0.5)$ find $\mathrm{P}(2 \leq X < 5)$.

```
> pnorm((4.5-6)/sqrt(3)) - pnorm((1.5-6)/sqrt(3))
[1] 0.1885507
> sum(dbinom(2:4,12,0.5))
[1] 0.1906738
```

Example (PQ, p80 Q21)**.** It is known that 80% of patients with a certain disease can be cured with a certain drug. What is the probability that amongst 150 patients with the disease, at most 37 of them cannot be cured with the drug.

Example. The proportion of children having a particular type of birth defect born to Pima Indian women is 0.05. Calculate the probability that in 785 independent births no more than 21 children have the birth defect.

Sampling distributions

- How do statistics vary across samples?
- Height for randomly selected $n = 4$ adult males.
- What is the distribution of $\overline{X}$ and S^2?

Model: Assume 4 independent readings of
Observations: X_1, X_2, X_3, X_4

$$\Rightarrow \overline{X} = \frac{1}{4}\sum_{i=1}^{4} X_i$$

The mean: because $\mathrm{E}\, X_i = 178$ and $\mathrm{Var}\, X_i = 8^2$ it follows $\overline{X} \sim \mathcal{N}(178, 4^2)$.
The sample variance: but $s^2 \not\sim \mathcal{N}$!

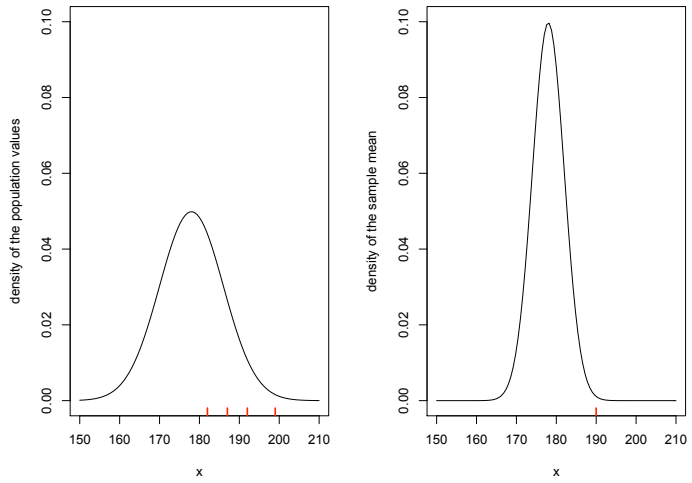
$$S^2 = \frac{1}{4-1}\Big((X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 + (X_4 - \overline{X})^2\Big).$$

Sampling distribution for S^2 and non-normal models

- Use CLT for large n and non-normal models.
- Knowing the sampling distribution helps identify unusual statistic values.
- E.g. if $\overline{X}$ was 190 (four basketball players):

```
# sampling distribution and extreme observations
x = c(182,187,192,199);
x.m = mean(x)
dnorm2 = function(x){
return(dnorm(x,mean=178,sd=8))}
dnorm3 = function(x){
return(dnorm(x,mean=178,sd=4))}
par(mfrow=c(1,2))
curve(dnorm2,from=150,to=210,ylim=c(0,0.1),ylab="density of the population values")
rug(x,col=2,lwd=2)
curve(dnorm3,from=150,to=210,ylim=c(0,0.1),ylab="density of the sample mean")
rug(x.m,col=2,lwd=2)
```

Distribution of population and mean



There must be something special with those 4 observations!

Sampling distributions – Movie 1

(Loading bootstrap.mp4)

Sampling distributions – Movie 2

(Loading bootstrap.mp4)

Lecture 3 - Content

- **Statistical inference**
- **Hypothesis testing**
- **One-sided tests for proportions**

Statistical inference

- Linking of observed data with possible statistical models or probability models.
- Based on some statistical model (i.e. assuming an underlying distribution, F, for observed data):
  - make decisions, e.g. in statistical hypothesis testing 'is the average measurement error equal to zero',
  - produce estimates, e.g. if the data is normal then use the mean to estimate the expected value,
  - make predictions, e.g. with time series, linear regression, and much more....

Random sample

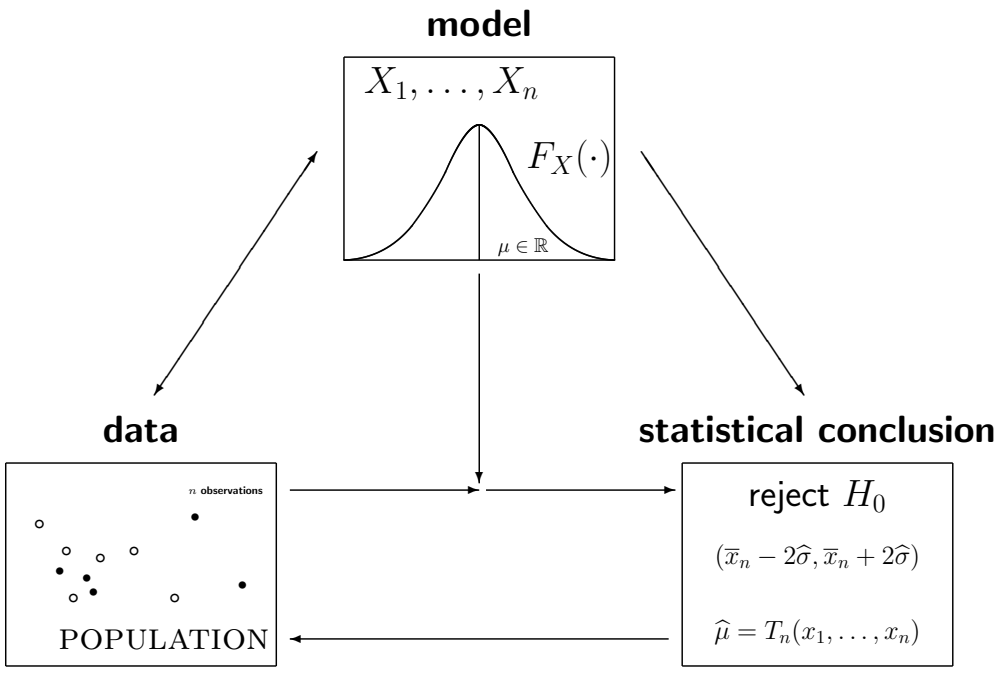
Statistical inference is inference about a population from a random sample drawn from it.

Definition 6. A set of observations (random variables) $X_1, \ldots, X_n$ constitutes a random sample of size n from the infinite population with cumulative distribution function $F(x) = \mathrm{P}(X \leq x)$ if:

- each X_i is a rv with identical CDF given by $F(x)$,
- these n random variables are independent.

Short notation: A sample $X_1, \ldots, X_n$ of length n is a set of n independent, identically distributed (iid) rvs with distribution F.

Statistical inference visualised



Three basic questions

1. Which parameter value serves based on the sample data as a *best guess* for an unknown model parameter?
   $\Rightarrow$ point estimation
2. Is there enough evidence based on the sample data to reject a pre-specified parameter value?
   $\Rightarrow$ hypothesis testing
3. Which possible parameter values of the statistical model are compatible with the sample data?
   $\Rightarrow$ interval estimation or confidence intervals

Hypothesis testing

Definition 7. A hypothesis, H, is a statement about an unknown parameter (e.g. μ) of the population.

This definition is vague by design.

Just about any kind of statement can count as a hypothesis, provided it is about a population parameter.

Hypothesis testing is the process of making a decision about a population parameter on the basis of statistics of an observed sample.

Definition 8. A null hypothesis, H_0, is a hypothesis set up to be nullified or refuted in order to support an alternative hypothesis, H_1.

In general the hypothesis test decides between two complementary hypotheses, H_0 and H_1. For example,

- H_0 may be a statement that the drug has no effect on controlling blood pressure and
- H_1 can be a statement that the drug has some effect on controlling blood pressure.

Typically H_0 is the simpler hypothesis, in the sense that it is about a parameter taking a specific value (rather than a range of values).

In hypothesis testing, one must decide either to accept H_0 as true or to reject H_0 as false and decide if H_1 is more plausible after observing the sample.

Definition 9. The critical region describes

- conditions under which H_0 should be rejected and
- conditions under which H_0 should be accepted.

General strategy:

- Find some statistic, τ (some function of our observed data).
- Find the distribution of τ assuming H_0 is true (called the null distribution).
- Calculate a corresponding P-value (defined below)
- Use the P-value to assess if data are consistent with H_0.

Definition 10. The P-value is the probability of getting an observed value of the test statistic or a more *unusual* value of the test statistic, under the assumption that H_0 is true.

Example

Most of these ideas can be illustrated by considering a coin toss example.

Let p, a parameter, be the probability of a head.

Assume the coin is 'fair so that at each toss we assume that $p = 0.5$. We call this the null hypothesis so that

$$H_0 \colon p = 0.5$$

and look for evidence against the null hypothesis H_0.

The only sensible alternatives are that:

- The coin is biased towards 'tails' in which case

$$H_1 : p < 0.5$$

- or the coin is biased towards 'heads' in which case $H_1 : p > 0.5$.

We look for evidence in favour of one of the alternatives by tossing the coin, say, 20 times and determine which of the hypotheses are most likely.

Example. Let X be the number of heads in 20 throws. Suppose we see 15 heads. Is the coin fair?

If the coin toss is fair then

$$X \sim \mathcal{B}(20, 0.5)$$

What is the chance of seeing exactly 15 heads?

$$\mathrm{P}(X = 15) = \text{dbinom(15,20,0.5)} = 0.01478577 \qquad \text{(which is small)}$$

(For continuous random variables analogous probabilities are zero, which is why we look for values of our test statistic as extreme or more extreme than what we observe).

What is the chance of seeing 15 heads or more?

$$P(X \geq 15) = \text{1 - pbinom(14,20,0.5)} = 0.02$$

which is still unlikely. Hence, H_0 is false or H_0 is true but we observed an unlikely outcome.

Example (Vaccination)**.** A flu vaccine is known to be 25% effective in the second year after inoculation. To determine if a new vaccine is more effective, 20 people are chosen at random and inoculated. If 9 of those receiving the new vaccine do not contract the virus in the second year after vaccination is the new vaccine superior to the old one?

- ☐
- ☐
- ☐
- ☐

- ☐

- ☐

Interpreting P-values

Uncertainty in the results: Because observations vary from sample to sample we can never say for sure whether H_0 is true or not.
Interpretation:

- Small P-values, for example a P-value of 0.01, means either
  - H_0 is true and the observed sample is improbable.
  - H_0 is not true.
- Large p-values, for example a P-value of 0.99 means either
  - the observed sample is consistent with H_0.
  - the observed sample comes from H_1, but by chance we are fooled into thinking the data comes from H_0.

The smaller the P-value, the stronger the evidence against H_0 in favour of H_1.

Some comments on the P-value

- If the P-value is small enough then we have evidence against H_0 in favour of the alternative hypothesis H_1.
- In the vaccination example we would conclude that the new vaccine is better.

- How small does the P-value have to be to decide in favour of H_1?
- There is no set value but

$$P\text{-value} \leq \alpha = 0.05 = 1/20$$

  is often used in practice. Other choices are: 0.1, 0.01, or 0.001 according to the 'innocent until proven guilty' principle.
- Under H_0, P-values have a uniform distribution or come very close to being uniform distributed!

Checklist for statistical tests

1. Hypotheses:
   - ☐ Null hypothesis, H_0.
   - ☐ Alternative hypothesis, H_1.
2. What is the test statistic, τ, and its sampling distribution if H_0 is true.
3. What is the critical region of the test statistic, i.e. which values of τ argue against H_0?
4. Observed test statistic (value of τ from the sample) and corresponding P-value.
5. Findings. If the P-value is small then either
   - ☐ H_0 is true and we have observed an unlikely event or
   - ☐ H_0 is false.

One-sided tests for proportions

Consider tests of

$$H_0 : p = p_0$$

against alternatives of the form

$$H_1 : p > p_0 \quad \text{or} \quad H_1 : p < p_0$$

for the distribution family $\mathcal{B}(n, p)$.

This situation occurs, say for example, when trying to determine (statistically) whether or not a coin is biased towards heads or tails.

Example

Example (Accid. Anal. and Prev. 1995:143-150)**.** A random sample of 319 front seat occupants involved in head-on collisions resulted in 95 who sustained no injuries. Does this support the claim that the proportion of uninjured occupants exceeds $1/3$?

Let $X =$ 'number of uninjured' in the sample and let

$$X \sim \mathcal{B}(319, p).$$

We wish to test $H_0 : p = 1/3$ against $H_1 : p > 1/3$.

Large values of X (our test statistic) argue for H_1.

Therefore the critical region will be the widest interval $[c_\alpha, \infty)$ such that

$$\mathrm{P}_{H_0}(X \geq c_\alpha) \leq \alpha.$$

The P-value is $\mathrm{P}(X \geq 95)$ calculated assuming H_0 is true.

Example (continued).

- In R with `1-pbinom(94,319,1/3)` or

```
> binom.test(95,319,1/3,alt="greater")
Exact binomial test
data:  95 and 319
number of successes = 95, number of trials =
319, p-value = 0.9211
alternative hypothesis: true probability of success is greater than 0.3333333
95 percent confidence interval:
 0.255656 1.000000
sample estimates:
probability of success
             0.2978056
```

Example (continued).

- or using the CLT: under $H_0 : X \simeq Y \sim \mathcal{N}(np, np(1-p))$, i.e. the

$$\begin{aligned} P\text{-value} &= \mathrm{P}(X \geq 95) = 1 - \mathrm{P}(X \leq 94) \\ &\simeq 1 - \mathrm{P}\left(Z \leq \frac{94.5 - 106.33}{\sqrt{70.89}}\right) \\ &= 1 - \Phi(-1.41) = 0.92 \text{ with } \texttt{1-pnorm(-1.405454)} \end{aligned}$$

$\Rightarrow$ there exists not enough evidence to support the claim that $p > 1/3$ but there is for any $p_0 \leq 0.253$.

```
> prop.test(95,319,1/3,alt="greater")
1-sample proportions test with continuity correction
data:  95 out of 319, null probability 1/3
X-squared = 1.6556, df = 1, p-value = 0.9009
alternative hypothesis: true p is greater than 0.3333333
95 percent confidence interval: 0.2560441 1.0000000
sample estimates:    p
             0.2978056
[P-value is different because there are various ways of correcting for continuity.]
```


R code

The code demonstrates the how P-values are uniformly distributed.

```
> set.seed(1)
> B = 10000  # no simulation runs
> n = 319    # sample size
> p = 1/3    # parameter value under Ho
> tau = rbinom(B,n,p)
> pvalue = 1 - pbinom( tau - 1 , n , p )  # alternative is p > 1/3
> hist(pvalue,breaks = 10)
```

Tuesday, 18 September 2012

Lecture 4 - Content

- ☐ **Two-sided tests for proportions**
- ☐ **Sign test**

Checklist for statistical tests

1. Hypotheses:
   - ☐ Null hypothesis, H_0.
   - ☐ Alternative hypothesis, H_1.
2. What is the test statistic, τ, and its sampling distribution if H_0 is true.
3. What is the critical region of the test statistic, i.e. which values of τ argue against H_0?
4. Observed test statistic (value of τ from the sample) and corresponding P-value.
5. Findings. If the P-value is small then either
   - ☐ H_0 is true and we have observed an unlikely event or
   - ☐ H_0 is false.

Two sided tests

Previously we only looked for alternatives of the form

$$H_1 : p > p_0 \quad \text{or} \quad H_1 : p < p_0.$$

These are called one-sided tests because they only consider the parameter lying to one side of a hypothesised value, in this case p_0.

In general we may not know in advance which alternative to choose. in this case we need to consider the two-sided hypothesis

$$H_1 : p \neq p_0$$

and in some cases this may be the only feasible alternative hypothesis.

WARNING: It is a statistical no-no to choose H_1 based on observed data. Instead H_1 should be chosen to dispel some preconceived outcome or alternatively based on expert opinion.

Test for proportions

Consider the two-sided hypothesis

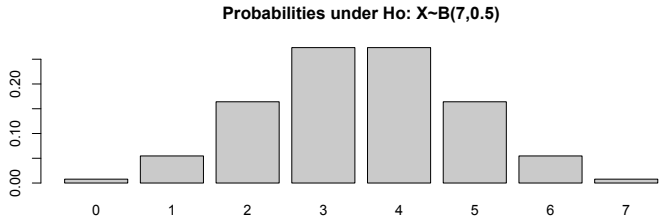
$$H_0 \colon p = p_0$$

where the general alternative is

$$H_1 \colon p \neq p_0.$$

Here we observe $X \sim \mathcal{B}(n, p)$, with $X \sim \mathcal{B}(n, p_0)$ under H_0:

$$\Rightarrow \text{large values of} \quad |X - np_0| \quad \text{argue against } H_0.$$



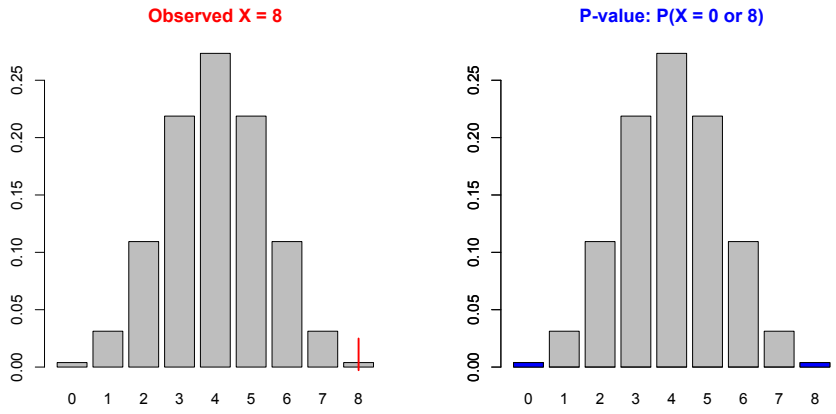
Example (Paul the octopus)**.** Is Paul the octopus guessing?

(http://en.wikipedia.org/wiki/Paul_the_octopus)

Paul correctly predicts 8 out of 8 winners in the 2010 World Cup!

- Let p denote the probability of correctly predicting the winner.
- Test: $H_0 : p = \frac{1}{2}$ against $H_1 : p \neq \frac{1}{2}$.
- Results: 8 of 8 winners in the 2010 World Cup were correctly predicted!
- Does this provide sufficient evidence against H_0?
- Test statistic: $X =$ 'no of correctly predicted winners in a sample of size n = '.
- Under H_0: $X \sim \mathcal{B}(8, 0.5)$; note $8 \times 0.5 < 5$, i.e. not yet with CLT.
- P-value: the values $X = 0$ and $X = 8$ are equally extreme or more extreme outcomes than the observed value of $X = 8$.

Example (cont).



- $\mathrm{P}(X \leq 0) + \mathrm{P}(X \geq 8) =$ `2*pbinom(0,8,0.5)` $=$ 0.0078125.
- Conclusion:

- Or much faster with `binom.test(8,8,1/2,alt="two.sided")`.

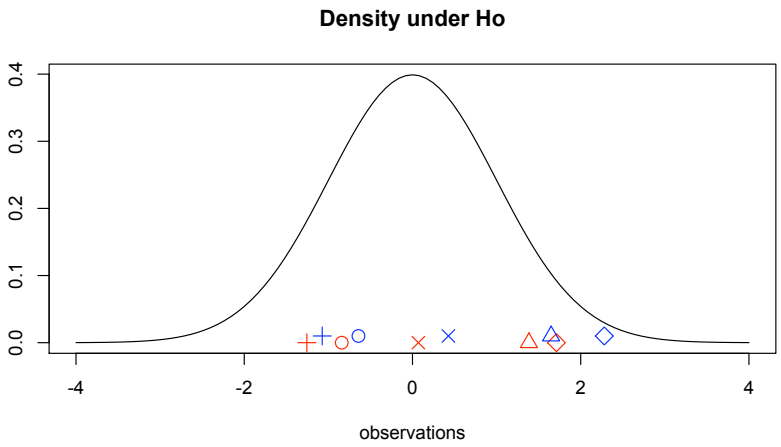
Example. A company claims that 93% of all items produced are non-defective. A random sample of 100 items is taken. If the observed number of defectives in the sample was 11 is there any reason to doubt the 93% claim?

- □
- □ Test
- □ Under H_0
- □ P-value

```
> 2*(1-pnorm(1.37))
[1] 0.1706869
> prop.test(11,100,0.07,alt="two.sided")
[...edited output...]
p-value = 0.1701
95 percent confidence interval: 0.05886717 0.19223346
```


Sign test

Paired data are very common. For example before/after trials, studies on twins, left/right arm freckles count.
Are the two samples from populations with the same distribution?



Analyse differences!

Theorem 6. If X and Y are iid with distribution function F then the distribution of $D = X - Y$ is symmetric with symmetry centre 0, i.e. $\mathrm{P}(D \leq -d) = \mathrm{P}(D \geq d)$ for all $d \in \mathbb{R}$.

Proof.

□

Constructing a simple test...

- Base a test on the number of positive differences.
- Hence, use the sign of the differences and ignore their magnitude

  $\Rightarrow$ test reduces to simple test of proportions.

Note, the simple test of proportions is for data with two possible outcomes only (yes/no, S/F, etc). Thus, we will discard differences which are exactly zero.

Example (Rats). A biochemical substance is believed to have an inhibitive effect on muscular growth. Ten laboratory rats of similar types are selected. For each rat

- one hind leg was regularly injected with the biochemical substance.
- The corresponding muscle on the other hind leg was regularly injected with a harmless placebo.
- At the end of 6 months the weights of the muscles were measured (in gms) and recorded as follows:

Rat	1	2	3	4	5	6	7	8	9	10
Bioch.	1.7	2.0	1.7	1.5	1.6	2.4	2.3	2.4	2.4	2.6
Placebo	2.1	1.8	2.2	2.2	1.5	2.9	2.9	2.4	2.6	2.5

- Analyse the data to determine whether this experiment provides evidence of a significant inhibitive effect.
- Why is this a good design for the study?

Example (cont).

- □
- □
- □
- □
- □
- □
- □
- □

Example (cont).

```
> # rat example
> x = c(1.7, 2.0, 1.7, 1.5, 1.6, 2.4, 2.3, 2.4, 2.4, 2.6)
> y = c(2.1, 1.8, 2.2, 2.2, 1.5, 2.9, 2.9, 2.4, 2.6, 2.5)
> d = y-x
> d
> plot(x,y,xlim=c(1.5,3),ylim=c(1.5,3))
> abline(0,1)
> text(2.75,1.5,"negative differences")
> text(1.75,3,"positive differences")
> points(c(1.8,2),c(1.8,1.8),type="l",lty=2,col="red")
> text(1.9,1.7,"y-x = -0.2")
> s = sign(d)[sign(d)!=0]
> table(s)
> binom.test(table(s),p=0.5,alt="less")
```

Example (Paint). A paint supplier claims that a new additive will reduce the drying time of acrylic paint. To test this claim 10 panels of wood are painted: one half with the original paint formula and one half with the paint having the new additive. The drying times in hours are given below.

```
> panel  = 1:10
> npaint = c(6.4,5.8,7.4,5.5,6.3,7.8,8.6,8.2,7.0,4.9)
> rpaint = c(6.6,5.9,7.8,5.7,6.0,8.4,8.8,8.4,7.3,5.8)
> d = rpaint - npaint
> d
 [1]  0.2  0.1  0.4  0.2 -0.3  0.6  0.2  0.2  0.3  0.9
```

- Can we conclude that the new additive is effective in reducing the drying time of the paint?
- Same steps as in previous example... but P-value $= 0.0107$.

Example (cont).

- The sign test can be used to test the hypothesis that the differences are scattered around 0.
- If the differences have a distribution that is symmetric about 0 then the probability of getting a positive difference, p_+, is 0.5.
- There are 10 non-zero differences.
- Test $H_0 : p_+ = \frac{1}{2}$ against $H_1 : p_+ > \frac{1}{2}$.
- Let X denote the number of positive differences. Large values of X support H_1. There are $m = 10$ non-zero differences. Thus if H_0 is true then $X \sim \mathcal{B}(10, 0.5)$.
- We observe 9 positive differences out of the $m = 10$ non-zero ones. P-value $= P(X \geq 9) = 1 - P(X \leq 8) = 1 - 0.9893 = 0.0107$. Since P is small we conclude that the new additive is effective in reducing the drying time of the paint.

Remarks

- Note the sign test ignores a lot of the information in the sample but it can be applied in quite general situations.
- Does not depend on the distribution of the data! For this reason sometimes these types of tests are called non-parametric.
- The sign test can be used to test if a single sample is taken from a continuous distribution that is symmetric about its population mean μ.

Monday, 1 October 2012

Lecture 5 - Content

☐ No lecture due to Labour Day holiday

Tuesday, 2 October 2012

Lecture 6 - Content

- **Tests for the mean μ**
- **Z-tests**

Reminder of Binomial/Sign Tests

For binomial/sign tests we have $\tau = X \sim \mathcal{B}(n, p)$.

For some fixed and known value p_0, or null hypothesis is

$$H_0 \colon p = p_0.$$

Under the assumption of H_0 we have $\tau = X \sim \mathcal{B}(n, p_0)$. We test H_0 against one of the following alternative hypotheses (with P-values),

$$H_1 \colon \begin{cases} p < p_0 & P\text{-value} = \mathrm{P}(X \leq x) \\ p > p_0 & P\text{-value} = \mathrm{P}(X \geq x) \\ p \neq p_0 & P\text{-value} = \mathrm{P}(|X - np_0| \geq |x - np_0|) \end{cases}$$

Reminder of P-values

Reminder: under H_0 the P-value is approximately $\mathcal{U}(0,1)$.

If the P-value is less than or equal to α (usually 5%) reject H_0. State there is statistical evidence against H_0 in favour of H_1.

If the P-value is greater than α accept H_0. State there is not sufficient statistical evidence to refute H_0 or the data is consistent with H_0. (DO NOT SAY THAT H_0 IS TRUE!!!).

Tests for the mean μ

Statistical tests can be developed to test claims about the population mean.

Assumption 0: Identically Distributed Since we are drawing samples from a particular population we implicitly assume that the samples are drawn from the same population, i.e. samples are identically distributed.

Assumption 1: Independence Assume that samples drawn from the population are selected independently, i.e. draws from the population do not depend on previous selections from the population

Assumption 2: Normal Samples (Stronger than Assumption 0) The population we are interested in has a Normal distribution, $\mathcal{N}(\mu, \sigma^2)$.

Tests for the mean μ

Suppose we have independent $X_1, \ldots, X_n$ with

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

An obvious test statistic to use for making inference about the mean μ is $\tau = \overline{X}$, the sample mean.

Two scenarios

At this point it is important to distinguish between two situations

- σ is known (e.g. IQ-test)
- σ is unknown, which is in general the case.

The distribution of $\tau = \overline{X}$ depends on whether σ is known or whether σ is unknown and needs to be estimated in some way.

Assumption 3: σ is known

The Z-test is constructed under the assumption that σ is known.

If the population variance, σ^2, is known the sampling distribution of the sample average is also known based on results stated in previous lectures.

If σ is known then the distribution of $\overline{X}$ is

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

where n is the sample size.

One-sided Z-test

- Test $H_0 \colon \mu = \mu_0$ against $H_1 \colon \mu > \mu_0$, where μ_0 is a given value.
- If H_0 is true then $\mu = \mu_0$ and so

$$\overline{X} \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right).$$

- Large values of $\overline{X}$ argue for H_1 (and against H_0).
- If the observed sample average is $\overline{x}$ the P-value is

$$P\text{-value} = \mathrm{P}(\overline{X} \geq \overline{x}) = \mathrm{P}\left(Z \geq \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right), \quad \text{where} \quad Z \sim \mathcal{N}(0,1).$$

Definition 11. The Z-value is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

and its corresponding test is called the Z-test.

Normal distributed data and n small

Example (Birthweights). The birthweights of a random sample of $n = 14$ boys born to mothers who smoked heavily during pregnancy were recorded (in ounces). The data are:

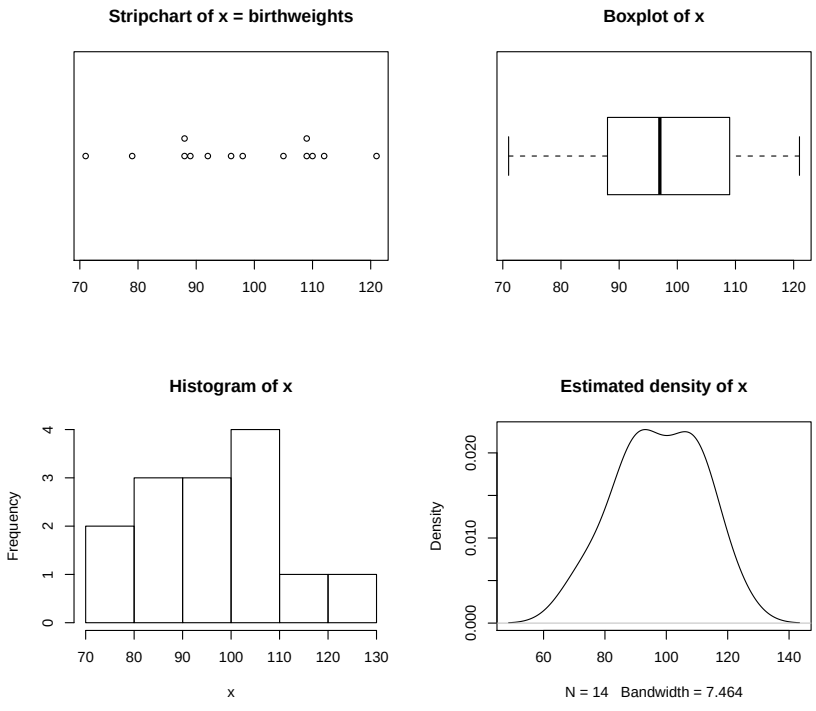
$$79,\ 92,\ 88,\ 98,\ 109,\ 109,\ 112,$$
$$88,\ 105,\ 89,\ 121,\ 71,\ 110,\ 96.$$

- It is believed that on average, boys born to mothers who smoke have a lower birthweight than the national average of 109 ounces (3.09kg).
- Is it reasonable to assume that birthweight has a normal distribution?
- Use R to explore ...

Example (cont)

```
> x = c(79,92,88,98,109,109,112,88,105,89,121,71,110,96)
> par(mfrow=c(2,2))
> stripchart(x, method="stack",offset=1, pch=1)
> title(main="Stripchart of x = birthweights")
> boxplot(x,range=1,horizontal=TRUE)
> title(main="Boxplot of x")
> hist(x)
> plot(density(x),main="Estimated density of x")
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  71.00   88.25   97.00   97.64  109.00  121.00
> IQR(x)
[1] 20.75
> sd(x)
[1] 14.05816
```


Example (cont)



Example (cont)

- Hence, we assume that the population of birthweights for boys born to mothers who smoke is modelled by
$$W \sim \mathcal{N}(\mu, 15^2).$$
- Test $H_0 \colon \mu = 109$ against $H_1 \colon \mu < 109$.
- The sample size is $n = 14$.
- Small values of $\overline{W}$ support H_1.
- If H_0 is true then the sampling distribution of $\overline{W}$ is
$$\overline{W} \sim \mathcal{N}\left(109, \frac{15^2}{14}\right).$$
- The observed value is $\overline{w} = \overline{x} = 97.64$ and $s = 14.05816$.

Example (cont)

-

- strong evidence against H_0.

Sample size n is large, normal or non-normal data

Example (SIDS victims)**.** In a random sample of 128 arterioles taken from SIDS (sudden infant death syndrome) victims the mean muscle thickness as a percentage of total arteriole diameter was 9.10.

- Assume that percentage muscle thickness can be modelled by

$$X \sim \mathcal{N}(\mu,\ 2.15^2).$$

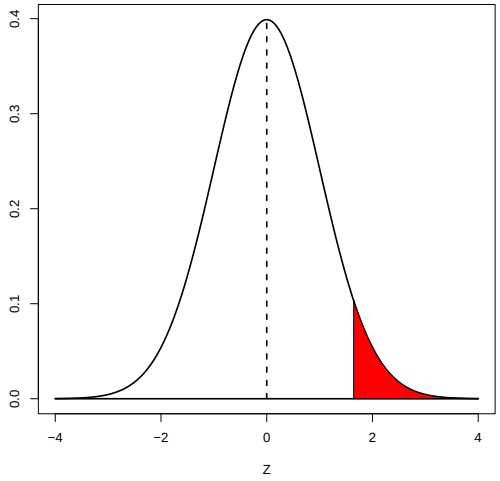
- For normal children of the same age $\mu = 6.04$.
- Is there evidence that the muscle thickness is greater in SIDS victims?

Example (cont)

- Test $H_0 \colon \mu = 6.04$ against $H_1 \colon \mu > 6.04$.
- Base the test on $\overline{X}$,

$$\overline{X} \sim \mathcal{N}(6.04, 2.15^2/128) \text{ if } H_0 \text{ is true.}$$

- Large values of $\overline{X}$ support H_1.



$$P\text{-value} = \mathrm{P}(\overline{X} \geq 9.10) = \mathrm{P}\left(Z \geq \frac{9.10 - 6.04}{2.15/\sqrt{128}}\right) = \mathrm{P}(Z \geq 16.10) < 10^{-4}$$

- Thus, the P-value is *very* small and so there is strong evidence against H_0.

Conclusions

- In the previous example the sample size was very large ($n = 128$).
- In such cases we know that the Central Limit Theorem (CLT) states that

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{(approx.)}$$

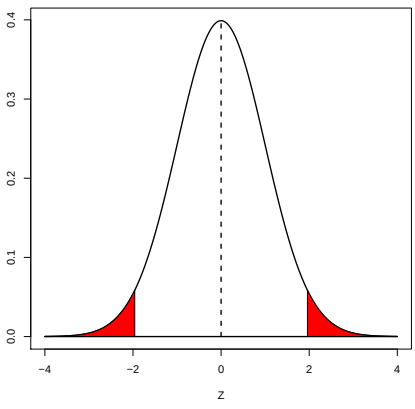
  whether the population is normal or not.
- Thus if the sample size is large then the CLT will enable us to calculate approximate P-values for tests of hypotheses about the mean regardless of the distribution of the underlying population provided σ is known.

Two-sided Z-tests

Example (Breaking strengths). A new synthetic fishing line is marketed with a manufacturer's claim that the mean breaking strength is 8 kgs with an s.d. of 0.5 kgs. Test this claim if a random sample of 50 lines is tested and the average of the sample of breaking strengths is $\overline{x} = 7.85$ kg.

- Here we have no reason to assume the true mean breaking strength is above or below 8 kgs if the claim is not true.
- Assume that the breaking strength can be modelled by $X \sim \mathcal{N}(\mu, 0.5^2)$.
- Test $H_0 \colon \mu = 8$ against $H_1 \colon \mu \neq 8$.
-
-

-



□

Conclusions from the previous three examples

- In all of the above examples we have been given the value for the population standard deviation, σ.
- In practice σ is generally unknown.
- In these cases how do we proceed?
- Recall the Z-test statistic is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

- We can estimate σ by using the sample standard deviation,

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

Monday, 8 October 2012

Lecture 7 - Content

- **One-sample t-tests**

One sample t-test

- In all examples in the last lecture(s) we were given the value for the population standard deviation σ,
- In practice σ is generally unknown!
- Estimate σ^2 by the sample variance s^2,

$$s = \sqrt{\frac{S_{xx}}{n-1}}$$

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \sum x_i^2 - n(\overline{x}^2). \end{aligned}$$

Theorem 7. If $\overline{X}$ is the mean of a sample of size n taken from a normal distribution having the mean μ and the variance σ^2, then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \text{ is a random variable}$$

having the t distribution with $\nu = n - 1$ degrees of freedom.

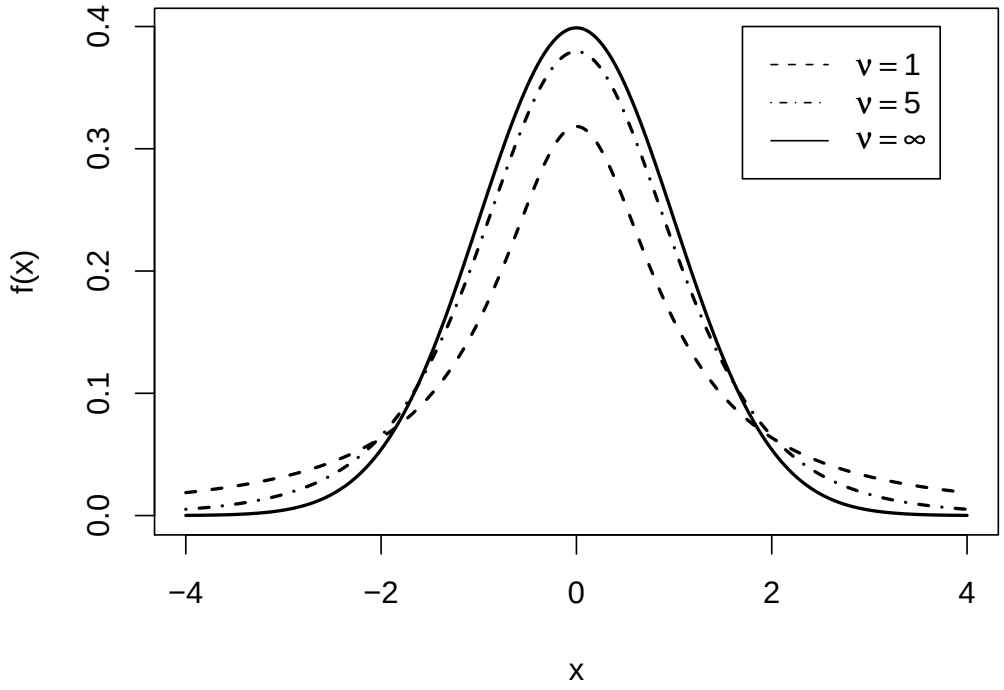
(Note that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.)

The t distribution

- The proof of the previous theorem will be shown in second year (need to show how to determine the distribution of a transformation of random variables).
- William S. Gosset (1908) (pen name: Student; statistician at Guinness)

- The density of the t distribution is symmetric and gets closer to the normal when $\nu = n - 1$ gets larger.
- Thicker tails of the t distribution takes into account the additional variability due to the estimation of σ by s.

The t distribution



The pdf of the t-distribution

Definition 12. A random variable having the t distribution with parameter $\nu = n - 1$ (degrees of freedom) has pdf (probability density function)

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

To say that the random variable T has the t distribution with $\nu \in \mathbb{N}$ df we write $T \sim t(\nu)$.

Remember: The Γ-function is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$$

and has the following properties (can be proved by partial integration):

$$\begin{aligned}\Gamma(\alpha+1) &= \alpha\Gamma(\alpha) \Rightarrow \Gamma(n+1) = n!;\ n \in \mathbb{N},\\ \Gamma(1/2) &= \sqrt{\pi}.\end{aligned}$$

Reminder of Assumptions

Assumption 0: Identically Distributed Since we are drawing samples from a particular population we implicitly assume that the samples are drawn from the same population, i.e. samples are identically distributed.

Assumption 1: Independence Assume that samples drawn from the population are selected independently, i.e. draws from the population do not depend on previous selections from the population

Assumption 2: Normal Samples (Stronger than Assumption 0) The population we are interested in has a Normal distribution, $\mathcal{N}(\mu, \sigma^2)$.

Z-tests assume that σ^2 is known.

Assumption 3: population is normal but σ^2 is unknown

- We can use a t-test when the population we are sampling from is normal but σ^2 is unknown.
- Check t-tables (formula sheet). Unlike the normal and binomial, t-tables are based on

$$\mathrm{P}(t_\nu > t) = p,$$

  where ν is the degree of freedom (row), p is the upper tail probability (column) and t is given in the body of the table.
- In R the following functions are helpful:
  - PDF: `dt(x,df=nu)`
  - CDF: `pt(q,df=nu)`
  - quantiles (critical values): `qt(p,df=nu)`
  - random numbers: `rt(n,df=nu)`

Example.

Example (Birthweights revisited)**.** Birthweights of boys to mothers who smoked:

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  71.00   88.25   97.00   97.64  109.00  121.00
> sd(x)
[1] 14.05816
```

- The 14 observations look like they could come from a normal distribution.
- Also, $\overline{w} = 97.643$ and $s = 14.058$.
- Test $H_0 \colon \mu = 109$ against $H_1 \colon \mu < 109$ using a t-test.
- Test statistic:

$$\tau = T = \frac{\overline{w} - 109}{s/\sqrt{14}},$$

  small values of τ support H_1.

Example (cont)

□

□

□

Example (cont)

```
> t.test(x,mu=109,alt="less")

One Sample t-test

data:  x
t = -3.0228, df = 13, p-value = 0.0049
alternative hypothesis: true mean is less than 109
95 percent confidence interval:
     -Inf 104.2966
sample estimates:
mean of x
 97.64286
```

One-sample t-tests continued

- Given a sample $X_1, \ldots, X_n$ from populations $\mathcal{N}(\mu, \sigma^2)$.
- Test $H_0 : \mu = \mu_0$ based on the test statistic

$$\tau = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}; \quad \text{where } s^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2 \text{ estimates } \sigma^2.$$

- If H_0 is true then,

$$\tau \sim t_\nu, \quad \text{where } \nu = \text{degrees of freedom.}$$

Example (Lubricants). The contents (in litres) of a random sample of 9 containers of lubricant are given:

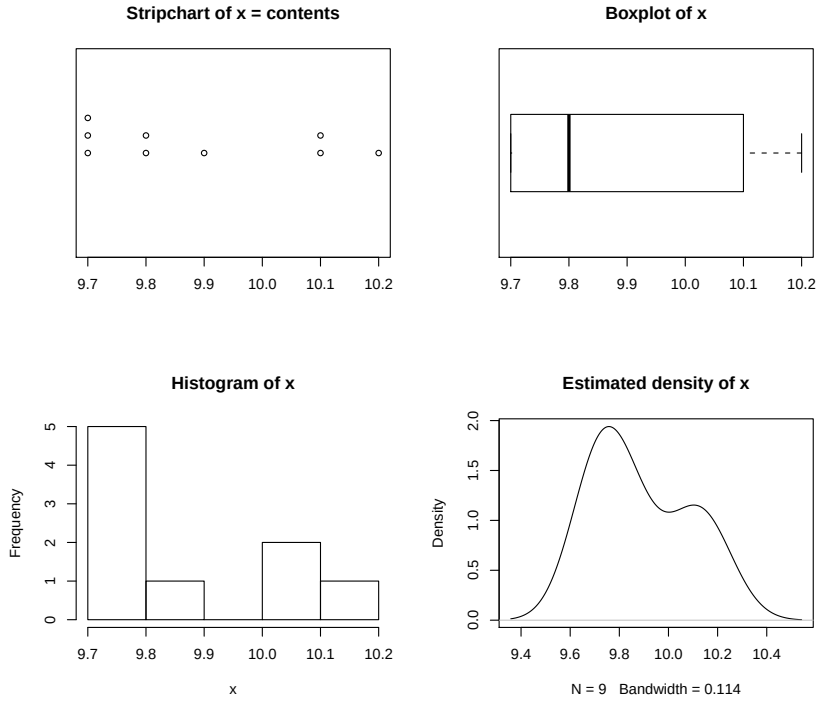
$$10.2,\ 9.7,\ 10.1,\ 9.7,\ 10.1,\ 9.8,\ 9.9,\ 9.8,\ 9.7.$$

Use these data to test the hypothesis that the (population) average content is 10 litres against the alternative that the true average contents is less than 10 litres.

- Can you assume that the contents $X \sim \mathcal{N}(\mu, \sigma^2)$? With R:

```
x = c(10.2,9.7,10.1,9.7,10.1,9.8,9.9,9.8,9.7)
stripchart(x, method="stack",offset=1, pch=1)
boxplot(x,range=1,horizontal=TRUE)
hist(x)
plot(density(x),main="Estimated density of x")
t.test(x,mu=10,alternative ="less")
```

Example (cont)



Example (cont)

- The sample average is $\overline{x} = 89/9 = 9.8889$.
- The sample standard deviation with R or by hand is

```
> sd(x) [...]    [1] 0.1964971
```

Tuesday, 9 October 2012

Lecture 8 - Content

- **One-sample t-tests continued**
- **Paired t-tests**

References from Phipps & Quine

- Section 3 pages 96–100.

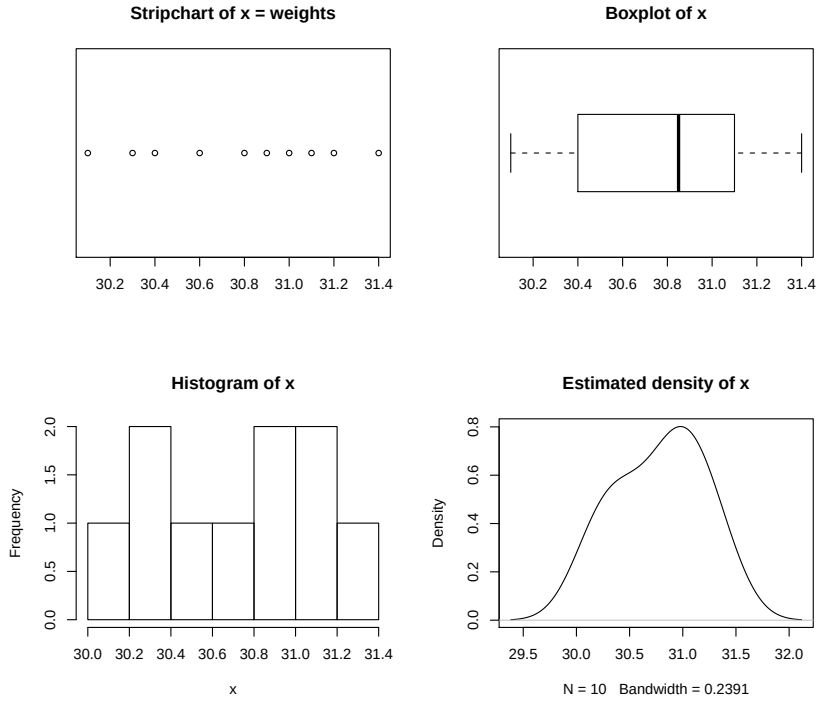
Example (Tablets). Ten tablets are weighed giving the weights (in mgs):

```
> x= c(31.0,31.4,30.4,30.1,30.6,31.1,31.2,30.9,30.3,30.8)
```

The machine producing these is set to give a mean weight of 30 mg. Is there evidence that the setting is incorrect?
Assume the weights are normally distributed $X \sim \mathcal{N}(\mu, \sigma^2)$.

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.10   30.45   30.85   30.78   31.08   31.40
> sd(x)
[1] 0.4211096
> stripchart(x, method="stack",offset=1, pch=1)
> boxplot(x,range=1,horizontal=TRUE)
> hist(x)
> plot(density(x))
```

Example (cont)



Example (cont)

- Sample size is small ($n = 10 < 25$), exploratory data analysis suggests normality may be reasonable (difficult to test for small sample sizes).
- The sample average is $\overline{x} = 30.78$.
- The sample standard deviation is $s = 0.4211096$.
- We wish to test,

$$H_0 \colon \mu = 30 \text{ against } H_1 \colon \mu \neq 30.$$

- Because sample size is very small, base the test on

$$\tau = \frac{\overline{X} - 30}{S/\sqrt{n}}.$$

- Either small values or large values of τ support H_1.

Example (cont)

- Under the assumption that the null hypothesis is true (along with independence and normality) the null distribution of the test statistic is $t_{n-1} = t_9$.

Example (cont)

- Alternatively, using the R commands,

```
> 2*pt(-5.857327,9)
[1] 0.000241544
> 2*pt(5.857327,9,lower.tail=F)
[1] 0.000241544
```

we get the exact P-value of 0.000241544.

Example (cont)

```
> t.test(x,mu=30,alternative ="two.sided")

One Sample t-test

data:  x
t = 5.8573, df = 9, p-value = 0.0002415
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 30.47876 31.08124
sample estimates:
mean of x
    30.78
```

Thus, there is strong evidence against H_0.

Summary of Z-tests and t-tests

- Assume the samples are independent and normally distributed.
- For some fixed and known value μ_0 the null hypothesis is $H_0\colon \mu = \mu_0$.
- If σ is unknown then $\tau = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$ and

$$H_1\colon \begin{cases} \mu < \mu_0 & P\text{-value} = \mathrm{P}(t_{n-1} \leq \tau) \\ \mu > \mu_0 & P\text{-value} = \mathrm{P}(t_{n-1} \geq \tau) \\ \mu \neq \mu_0 & P\text{-value} = 2\,\mathrm{P}(t_{n-1} \geq \tau) \end{cases}$$

- If σ is known then $\tau = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ or if σ is unknown and n is large ($n > 25$ so that the CLT applies) then $\tau = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \sim N(0,1)$ and

$$H_1\colon \begin{cases} \mu < \mu_0 & P\text{-value} = \mathrm{P}(Z \leq \tau) \\ \mu > \mu_0 & P\text{-value} = \mathrm{P}(Z \geq \tau) \\ \mu \neq \mu_0 & P\text{-value} = 2\,\mathrm{P}(Z \geq \tau) \end{cases}$$

Paired data

- Paired data are very common,
  - before/after trials
  - studies on twins
  - left arm vs right arm or left eye vs right eye experiments
- We can test if the two (paired) samples come from populations with the same mean by focusing on differences.
- Have differences zero mean?

Paired data - Assumptions

We have data of the form

X	X_1	X_2	$\ldots$	X_n
Y	Y_1	Y_2	$\ldots$	Y_n
D	D_1	D_2	$\ldots$	D_n

where $D_i = X_i - Y_i$. To perform a t-test we needed to assume

- Normality (hence identically distributed)
- Independence

where we do not know the variance of the data.

Paired data - Assumptions

- For the Paired t-test we assume that the differences $D_1, \ldots, D_n$ are independent normally distributed random variables.
- We not make assumptions on the Xs or Ys except that the Xs and Ys are *not independently obtained*, i.e. there is a natural pairing of the data.
- Later for the two-sample t-test (another test involving two sets of data) we assume that X and Y are *independently obtained*.
- The paired t-test is similar in spirit to the sign-test. However, for the sign-test we assume symmetry while for the paired t-test we make the stronger assumption that the differences are normally distributed.
- Also, the sign-test removes zero differences, whereas the paired t-test uses all available observations.

Example (Rats, PQ p125 and L16)**.** Does a biochemical substance have an inhibitive effect on muscular growth? For each of 10 rats:

- one hind leg was regularly injected with the biochemical substance.
- The corresponding muscle on the other hind leg was regularly injected with a harmless placebo.
- At the end of 6 months the weights of the muscles were measured (in gms) and recorded as follows:

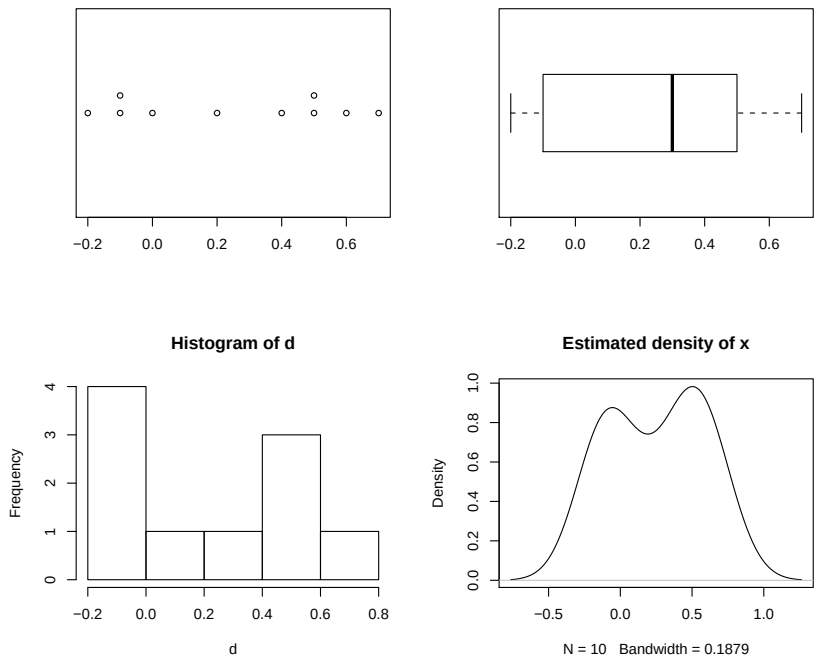
Rat	1	2	3	4	5	6	7	8	9	10
Bioch.	1.7	2.0	1.7	1.5	1.6	2.4	2.3	2.4	2.4	2.6
Placebo	2.1	1.8	2.2	2.2	1.5	2.9	2.9	2.4	2.6	2.5

- Analyse the data to determine whether this experiment provides evidence of a significant inhibitive effect.

Example (cont)

```
> x = c(1.7, 2.0, 1.7, 1.5, 1.6, 2.4, 2.3, 2.4, 2.4, 2.6)
> y = c(2.1, 1.8, 2.2, 2.2, 1.5, 2.9, 2.9, 2.4, 2.6, 2.5)
> d = y-x
> par(mfrow=c(2,2))
> stripchart(d, method="stack",offset=1, pch=1)
> boxplot(d,range=1,horizontal=TRUE)
> hist(d)
> plot(density(d),main="Estimated density of x")
```


Example (cont)



- Sample size is small ($n = 10 < 25$), exploratory data analysis suggests normality may be reasonable (again, difficult to test for small sample sizes).

```
> mean(d)
[1] 0.25
< sd(d)
[1] 0.3308239
```

- The sample average is $\overline{x} = 0.25$ and the sample standard deviation is $s = 0.3308239$.
- We wish to test,

$$H_0 \colon \mu_d = 0 \quad \text{against} \quad H_1 \colon \mu_d > 0.$$

- Again, because sample size is very small, base the test on

$$\tau = \frac{\overline{X}}{S/\sqrt{n}}.$$

- Either small values or large values of τ support H_1.

Example (cont)

- Under the assumption that the null hypothesis is true (along with independence and normality) the null distribution of the test statistic is $t_{n-1} = t_9$.
-
-
-

Example (cont)

- Alternatively, using the R commands,

```
> 1- pt(2.389699,9)
[1] 0.02028870
> pt(2.389699,9,lower.tail=F)
[1] 0.02028870
```

we get the exact P-value of 0.02028870.

Example (cont)

- Alternatively, using R:

```
> t.test(d,mu=0,alternative ="greater")

        One Sample t-test

data:  d
t = 2.3897, df = 9, p-value = 0.02029
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 0.05822761        Inf
sample estimates:
mean of x
     0.25
```

Example (cont)

- Via a sign test we obtain

```
> s = sign(d)[sign(d)!=0]
> table(s)
s
-1  1
 3  6
> binom.test(c(6,3),p=0.5,alt="greater")

        Exact binomial test

data:  c(6, 3)
number of successes = 6, number of trials = 9, p-value = 0.2539
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.3449414 1.0000000
sample estimates:
probability of success
             0.6666667
```

- In this case the t-test and the sign-test give conflicting results. This is not uncommon when the sample size is small.

Example – Paint

Example (Paint, continued from L16)**.** A paint supplier claims that a new additive will reduce the drying time of acrylic paint. To test this claim 10 panels of wood are painted: one half with the original paint formula and one half with the paint having the new additive. The drying times in hours are given below.

```
> panel  = 1:10
> npaint = c(6.4,5.8,7.4,5.5,6.3,7.8,8.6,8.2,7.0,4.9)
> rpaint = c(6.6,5.9,7.8,5.7,6.0,8.4,8.8,8.4,7.3,5.8)
> d = rpaint - npaint
> d
 [1]  0.2  0.1  0.4  0.2 -0.3  0.6  0.2  0.2  0.3  0.9
```

- Can we conclude that the new additive is effective in reducing the drying time of the paint?
- Same steps as in previous example.

Example (cont)

-
-
-

Monday, 15 October 2012

Lecture 9 - Content

- ☐ **Two-sample t-tests**
- ☐ **Confidence intervals**

Two-sample t-tests

Assumptions

Two independent samples with n_x observations $x_1, \ldots, x_{n_x}$ from one population and n_y observations $y_1, \ldots, y_{n_y}$ from another. We assume that the populations can be modelled by $\mathcal{N}(\mu_x, \sigma^2)$ and $\mathcal{N}(\mu_y, \sigma^2)$:

(i) Two independent samples from

(ii) normal populations with

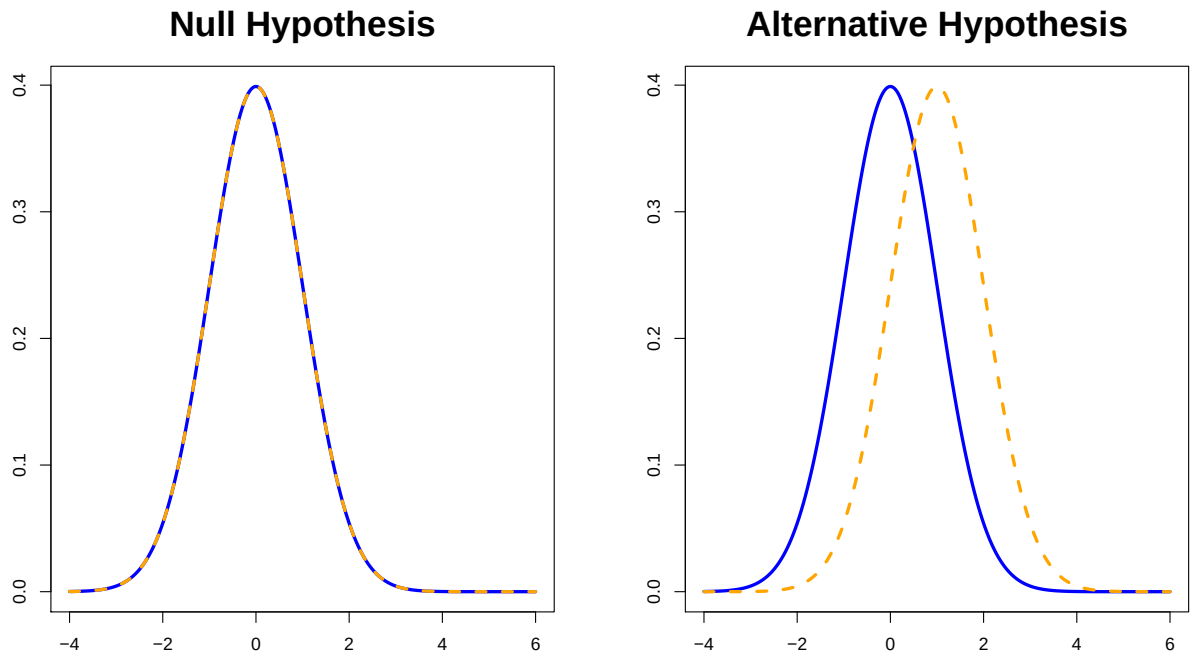
(iii) common variance.

Example (Height and gender: http://en.wikipedia.org/wiki/Human_height).

Country/Region	Average male height (m)	Average female height (m)	Age range	Method	Year
Argentina	1.735	1.608	17	Measured	1998-2001
Australia	1.748	1.635	18+	Measured	1995
Austria	1.796	1.671	21-25	Self Reported	1997-2002
Azerbaijan	1.718	1.654	16+	Measured	2005
Bahrain	1.651	1.542	19+	Measured	2002
Belgium	1.795	1.678	21-25	Self Reported	1997-2002
Bolivia	1.600	1.422	20-29	Measured	1970
Brazil	1.707	1.588	18+	Measured	2008-2009

Average height of Australians (to 0 d.p.): $\mu_x = 164$ and $\mu_y = 175$ with standard deviation typically in the range of $\sigma \in (6.5\text{cm}, 7.5\text{cm})$

Two Sample t-test



Testing equality of population means

How do we test

$$H_0 : \mu_x = \mu_y \quad \text{against} \quad H_1 : \mu_x \neq \mu_y \, ?$$

Available information:

- sample sizes: n_x and n_y
- sample means: $\overline{x}$ and $\overline{y}$
- sample variances: s_x^2 and s_y^2

Test statistic: If σ^2 is known then the differences of the means has distribution

$$\overline{X} - \overline{Y} \quad \overset{\text{if } H_0 \text{ is true}}{\Rightarrow} \quad \overline{X} - \overline{Y} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}\right)$$

Two-sample Z- and t-test statistics

Hence, if $H_0 : \mu_x = \mu_y$ is true and σ^2 is known,

$$\frac{\overline{X} - \overline{Y}}{\sigma\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} = Z \sim \mathcal{N}(0, 1)$$

and more generally, if σ^2 is unknown and can be estimated by the pooled variance

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)}$$

thus,

$$\tau = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t_{(n_x+n_y-2)},$$

i.e. a t-distribution with degrees of freedom equal $\nu = n - 2 = n_x + n_y - 2$.

Example (Height and gender: http://en.wikipedia.org/wiki/Human_height (cont)). Suppose that in a particular MATH1905 tutorial we have $\overline{x} = 164$, $\overline{y} = 175$, $s_x = 6.8$, $s_y = 7.2$, $n_x = 8$, $n_y = 9$ and we want to test whether males are taller than females in the MATH1905 tutorial.

- The null and alternative hypotheses are

$$H_0 \colon \mu_x = \mu_y \quad \text{versus} \quad H_1 \colon \mu_x < \mu_y.$$

- The pooled variance is given by

$$s_p^2 = \frac{(8-1) \times 6.8^2 + (9-1) \times 7.2^2}{(8+9-2)} = 49.22667.$$

- The observed value of the test statistic is

$$\tau = \frac{164 - 175}{\sqrt{49.22667} \times \sqrt{\frac{1}{8} + \frac{1}{9}}} = -3.226519$$

- Large (negative) values provide evidence for H_1.

- Assuming independent normal observations with common variance and under the null hypothesis the null distribution of the test statistic is $t_{(n_x+n_y-2)} = t_{15}$.
- The P-value for this hypothesis is given by

$$P(t_{15} < -3.226519) = \text{pt(-3.226519, 15)} = 0.002824303$$

- Hence, we reject the null hypothesis (that male and female heights in the MATH1905 tutorial are equal) in favour of the alternative hypothesis (that men and taller than women in the MATH1905 tutorial are equal).

Example (Fusion of Ice)**.** Two methods, A and B were used in the determination of the latent heat of fusion of ice. The investigators wished to find out whether the methods differed. The following table gives the change in total heat from ice at $-0.72°$C to water at $0°$C in calories per gram.

A: 79.98 80.04 80.02 80.04 80.03 80.03 80.04
80.05 80.03 80.02 80.00 80.02 79.97

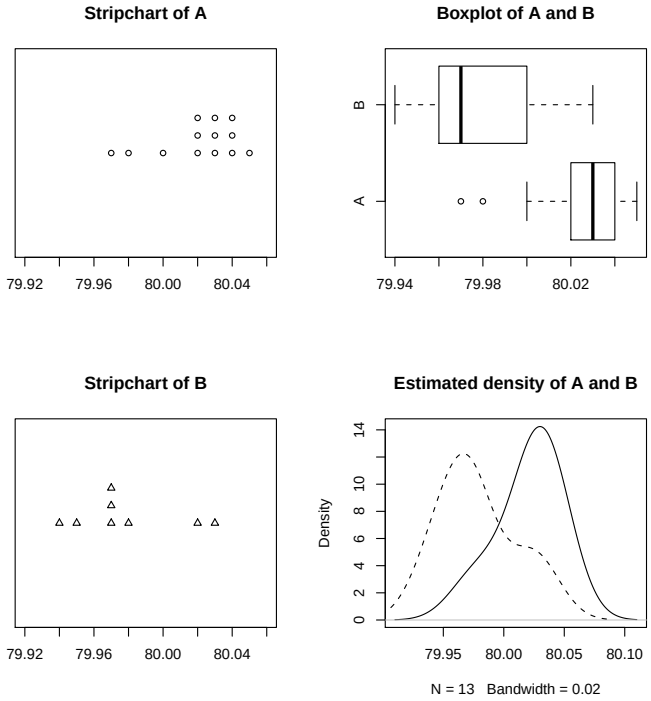
B: 80.02 79.94 79.98 79.97
79.97 80.03 79.95 79.97

- Assume the change in total heat values can be modelled in each case by a normal distribution.
- Do you agree?

Example (cont)

```
> A = c(79.98,80.04,80.02,80.04,80.03,80.03,80.04,80.05,
        80.03,80.02,80.00,80.02,79.97)
> B = c(80.02,79.94,79.98,79.97,79.97,80.03,79.95,79.97)
> par(mfrow=c(2,2))
> stripchart(A, method="stack",offset=1, pch=1,xlim=c(79.92,80.06))
> title(main="Stripchart of A")
> boxplot(c(A,B)~c(rep("A",13),rep("B",8)),range=1,horizontal=TRUE)
> title(main="Boxplot of A and B")
> stripchart(B, method="stack",offset=1, pch=2,xlim=c(79.92,80.06))
> title(main="Stripchart of B")
> plot(density(A,bw=0.02),main="Estimated density of A and B")
> points(density(B,bw=0.02),type="l",lty=2)
> c(length(A),length(B)) ... edited ...   [1] 13     8
> c(mean(A),mean(B)) ... edited ...       [1] 80.02 79.98
> c(sd(A),sd(B)) ... edited ...           [1] 0.024 0.031
```

Example (cont)



Example (cont)

A: $n_x = 13 \quad \overline{x} = 80.0208 \quad s_x = 0.02397$

B: $n_y = 8 \quad \overline{y} = 79.9788 \quad s_y = 0.03137$

- ☐
- ☐
- ☐
- ☐
- ☐
- ☐

In R with `pt()` command or by `t.test(A,B,mu=0,var.equal=TRUE)`.

```
        Two Sample t-test

data:  A and B
t = 3.4722, df = 19, p-value = 0.002551
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01669058 0.06734788
sample estimates:
mean of x mean of y
 80.02077  79.97875
```


Summary of Hypothesis Testing

1. Tests for Proportions: $X \sim \mathcal{B}(n, p)$

   $$H_0 : p = p_0$$

   Base the test on X and use binomial tables or the normal approx. to get the P-value of the **binomial test**.

2. Tests of the Mean - Single Sample:

   $$H_0 : \mu = \mu_0.$$

   (i) Population is symmetric
   Use the **sign test** which is based on the test for proportions and the number of positive signs with $p_0 = 0.5$.

2. Tests of the Mean - Single Sample (cont):

(ii) Population is $\mathcal{N}(\mu, \sigma^2)$ with σ *known*.
Use the **Z-test**

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

(iii) Population is $\mathcal{N}(\mu, \sigma^2)$ with σ *unknown* and n is small ($n < 25$).
Use the **t-test**.

$$\tau = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

(iv) Population is $\mathcal{N}(\mu, \sigma^2)$ with σ *unknown* and n is large ($n > 25$).
Then **Z-test** approx. **t-test**.

$$\tau = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim N(0,1)$$

3. Tests of Means - Two Samples:

Are the data paired?

(a) Yes - Calculate the differences.

(i) Differences have a symmetric distribution about μ
Use the **sign test** to test $H_0 : \mu = 0$.

(ii) Differences have a $\mathcal{N}(\mu, \sigma^2)$ distribution
Use the **t-test** to test $H_0 : \mu = 0$.

(b) No - Are the samples independent?
Are the populations Normal with common variance? If 'yes', use the

2 sample t-test to test $H_0 : \mu_x = \mu_y$

$$\tau = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}.$$

Confidence intervals

- Given a sample $X_1, \ldots, X_n$ from a normal population $X \sim \mathcal{N}(\mu, \sigma^2)$ how do we estimate μ?
- The best estimate in the least squares or maximum-likelihood sense is

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- $\overline{X}$ is close to the true μ but with probability one wrong, i.e.

$$\mathrm{P}(\overline{X} = \mu) = 0 \quad \text{since the normal is continuous.}$$

- Known result: If σ is known then,

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

  and thus $\mathrm{P}(-1.96 \leq Z \leq 1.96) = 0.95 \Rightarrow$ substitute Z and solve for μ.

95% CI for μ if σ is known

Thus,

95% CI for μ if σ is known

$$0.95 = \mathrm{P}\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

We can interpret this equation as saying:

> If we were to repeat the experiment over and over again (with the same sample size) and recalculate the confidence interval each time then 95% of the calculated confidence intervals will contain the true value of μ.

Using statistical jargon we say the random interval

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

covers μ with probability 0.95.

Another of John's pet hates

The wrong interpretation is:

> There is a 95% chance that the population mean is between 165cm and 189cm.

The correct interpretation is:

> For 95and 189cm covers the population mean.

Note that the "randomness" is on the fact that samples are drawn from a particular population, **not in the parameter of interest!**

$100(1-\alpha)\%$ CI for μ if σ is known

Definition 13. The $100(1-\alpha)\%$ CI for μ is given by

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

and is constructed by finding $z_{\alpha/2}$ such that

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

and solving for μ.

Example (Cholesterol Levels)**.** Consider the distribution of serum cholesterol levels for all males in the United States who are hypersensitive and who smoke. The distribution is normal with an unknown mean and a known variance of 46 mg/100 ml (based on historical records). Suppose that we draw a random sample of size $n = 12$ from the population of interest which has sample average $\overline{x} = 217$ mg/100 ml. What is the 95% confidence interval the population mean μ?

□

□

Tuesday, 16 October 2012

Lecture 10 - Content

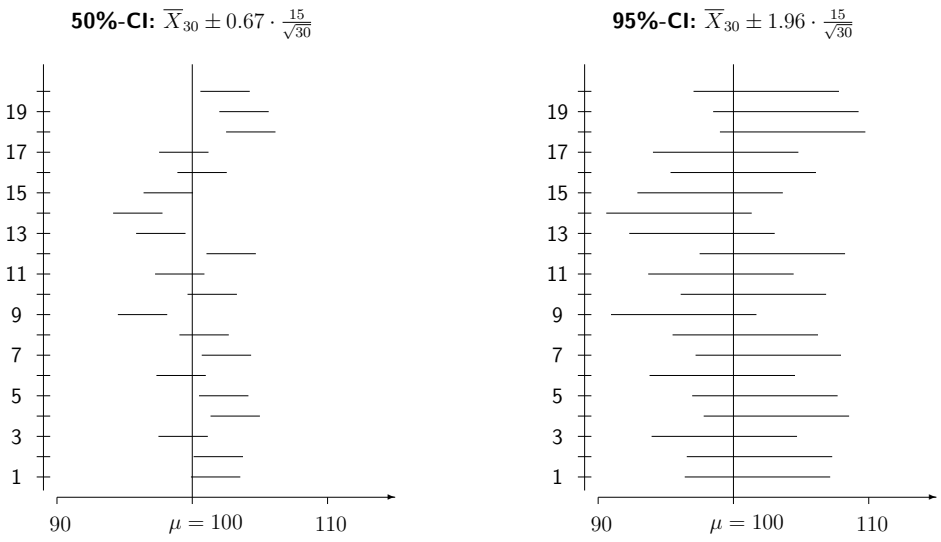
- **Confidence intervals continued**

Confidence intervals (cont)

Properties of CIs

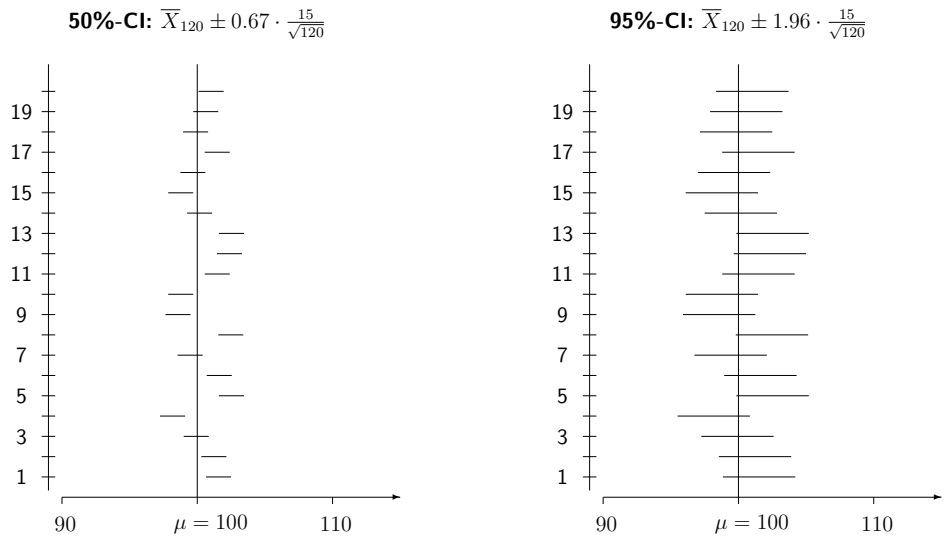
- Cover the true μ value with relative frequency approximately $(1-\alpha)$;
- as you increase n the CI gets narrower;
- as you increase the confidence level, i.e. make $(1-\alpha)$ larger, the CI gets wider.

Simulated CIs for IQ tests, $n = 30$:



Here, population variance is $\sigma^2 = 15^2$ and population mean $\mu = 100$.

Simulated CIs for IQ tests, $n = 120$:



Here, population variance is $\sigma^2 = 15^2$ and population mean $\mu = 100$.

Example (Birthweight). Use the following data to construct a 90% and 99% CI for the average birthweight of a term baby (37 - 41 weeks gestation) if it is known that the birthweight (in kgs) is $W \sim \mathcal{N}(\mu, 0.525^2)$.

$$2.853,\ 3.127,\ 3.159,\ 3.800,\ 2.656,\ 3.245,\ 3.510,\ 3.082$$

- $\overline{x} = 25.432/8 = 3.179$.
- 90% CI for μ: Find z such that $0.90 = \mathrm{P}(-z \leq Z \leq z)$, that is, $\mathrm{P}(Z > z) = 0.05$. From t-tables with $\nu = \infty$, z-tables or with R: $z = 1.645$.
- C.I. calculates to $3.179 \pm 1.645 \times \frac{0.525}{\sqrt{8}} = 3.179 \pm 0.305 = (2.874,\ 3.484)$.
- 99% C.I. for μ: $0.99 = \mathrm{P}(-z_1 \leq Z \leq z_1) \Rightarrow z_1 = 2.576$ and CI is $3.179 \pm 2.576 \times \frac{0.525}{\sqrt{8}} = 3.179 \pm 0.478 = (2.701,\ 3.657)$.

$100(1-\alpha)\%$ CI for μ if σ is unknown

Base the CI on the t-statistic,

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1},$$

where S^2 the sample variance.

Definition 14. A $100(1-\alpha)\%$ CI for μ of a normal population with unknown variance σ^2 is given by

$$\overline{X} \pm t' \frac{s}{\sqrt{n}},$$

where t' is from the t-tables or from R such that

$$1-\alpha = \mathrm{P}(-t' \leq t_{n-1} \leq t').$$

Example. Consider the distribution of serum cholesterol levels for all males in the United States who are hypersensitive and who smoke. The distribution is normal with an unknown mean and a unknown variance. Suppose that we draw a random sample of size $n = 12$ from the population of interest which has sample average of 217 mg/100 ml and sample variance of 46. What is the 95% confidence interval the population mean μ?

□

□

- Note that when we assumed $\sigma = s = \sqrt{46}$ we obtained the confidence interval

$$(213.16, 220.84)$$

- Notice the confidence intervals are slightly wider taking into account the uncertainty when estimating σ by s.

Example (Paint). The 10 values below are the first sample of values on paint primer thickness that were collected as part of an ongoing process of monitoring the performance of an industrial system.

$$1.30,\ 1.10,\ 1.20,\ 1.25,\ 1.05,$$

$$0.95,\ 1.10,\ 1.16,\ 1.37,\ 0.98$$

- Assume the primer thickness can be modelled by $X \sim \mathcal{N}(\mu, \sigma^2)$.
- $\overline{x} = 1.146, \quad s = 0.1363.$
- A 95% C.I. for μ is $\overline{x} \pm t' \frac{s}{\sqrt{10}}$, where $0.95 = \mathrm{P}(-t' \leq t_9 \leq t')$.
- $\mathrm{P}(t_9 > t') = 0.025$ thus, $t' = 2.262$.
- The CI is $1.146 \pm 2.262 \times \frac{0.1363}{\sqrt{10}} = 1.146 \pm 0.097$ or $(1.049,\ 1.243)$.

CIs for proportions

Data: n independent trials and the probability of success at each trial is p, X denotes the number of successes,

$$\Rightarrow X \sim \mathcal{B}(n,p), \quad \mathrm{E}(X) = np, \quad \mathrm{Var}(X) = np(1-p).$$

Standardized scores: Calculate standardized number of successes,

$$Z' = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}},$$

where $\widehat{p} = X/n$ is the sample proportion (estimated proportion).

- If n is large: $Z' \simeq \mathcal{N}(0,1) \Rightarrow$ use Z' to obtain approximate CIs for p.
- However, the variance depends also on the unknown parameter p!
- $\mathrm{Var}\, X/n = p(1-p)/n \approx \widehat{p}(1-\widehat{p})/n \leq \frac{1}{2}\left(1-\frac{1}{2}\right)/n = \frac{1}{4n}$.

Definition 15. An approximate $100(1-\alpha)\%$ CI for p is obtained from

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

and a conservative CI for p is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{1}{4n}}.$$

Example. What sample size is necessary to give a 95% C.I. for a proportion with width ± 0.03? [Note that as a convention, width is the same as the length of the CI, i.e. ± 0.03 corresponds to a length = width = 0.06]

-
- 1068

Example. A new type of photoflash bulb was tested to estimate the probability, p, of producing the required light output at the appropriate time. The sample of 1000 bulbs were tested and 810 were observed to function according to specifications. Estimate p and find and approximate 95% confidence interval for p.

- □
- □
- □

Comments on opinion polls

- ACNeilsen and others poll typically about $1,000$ people.
- Why?
- The conservative $\pm$ factor for a 95% C.I. is

$$1.96/\sqrt{4 \times 1000} = 0.031$$

  hence the margin of error is about 3 percent.
- As a rough guide the margin of error is

$$\frac{1.96}{\sqrt{4n}} \simeq \frac{1}{\sqrt{n}}.$$

Example (Sample sizes in surveys)**.** A survey is to be conducted to determine the proportion of a population with a certain attribute.

(i) What sample size is necessary to ensure the sample proportion is within 0.03 of the true population proportion with probability at least 0.9?

Solutions:

(i) We want n such that $\mathrm{P}(|\widehat{p} - p| < 0.03) \geq 0.90$. $\widehat{p}$ is approximately normally distributed with variance $p(1-p)/n$ so we want $P\left(|Z| < \frac{0.03 \times \sqrt{n}}{\sqrt{p(1-p)}}\right) \geq 0.90$

$$\frac{0.03 \times \sqrt{n}}{\sqrt{p(1-p)}} \geq 1.645 \overset{\text{solve for } n}{\Rightarrow} n \geq \left(\frac{1.645}{0.03}\right)^2 \times p(1-p).$$

If we replace $p(1-p)$ by $\frac{1}{4}$ then we have $n \geq 751.67$ so a sample of size 752 will certainly suffice.

Example (Sample sizes in surveys)**.** A survey is to be conducted to determine the proportion of a population with a certain attribute.

(ii) What sample size is needed so that a 95% C.I. has width no more than 0.04 (i.e. the $\pm$ term is less than 0.02)?

Solutions:

(ii) We use the conservative version of the C.I. and recall the 95% C.I. $\pm$ factor is always less than

$$1.96\sqrt{\frac{1}{4n}}.$$

Solve

$$\frac{1.96}{2\sqrt{n}} \leq 0.02 \Rightarrow \frac{1.96}{2 \times 0.02} \leq \sqrt{n} \Rightarrow 2401 \leq n.$$

Thus a sample of 2401 observations is needed.

Example (Sample sizes in surveys)**.** A survey is to be conducted to determine the proportion of a population with a certain attribute.

(iii) As in (ii) but assuming that the true proportion will be less than 30%?

Solutions:

(iii) Because $p \leq 0.3$ we get a smaller conservative bound of $\operatorname{Var} Z' \leq 0.3 \times 0.7/n$. Hence, for the 95% CI the $\pm$ factor is always less than

$$1.96\sqrt{\frac{0.21}{n}}.$$

Solve

$$\frac{1.96 \times \sqrt{0.21}}{\sqrt{n}} \leq 0.02 \;\Rightarrow\; \frac{1.96 \times \sqrt{0.21}}{0.02} \leq \sqrt{n} \;\Rightarrow\; 2016.84 \leq n.$$

Thus a sample of 2017 observations is needed.

Summary of Confidence Interval

We have covered the following cases:

- Normal/Constant $\sigma^2 = \sigma_0^2$ case: $\overline{x} \pm z^* \times \frac{\sigma_0}{\sqrt{n}}$
- Normal/Unknown $\sigma^2/n < 30$ case: $\overline{x} \pm t^* \times \frac{s}{\sqrt{n}}$
- Normal/Unknown $\sigma^2/n \geq 30$ case: $\overline{x} \pm z^* \times \frac{s}{\sqrt{n}}$
- Proportions: $\widehat{p} \pm z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$
- Proportions (Conservative): $\widehat{p} \pm z^* \sqrt{\frac{1}{4n}}$

where $P(|Z| \leq z^*) = 1 - \alpha$, $P(|t_{n-1}| \leq t^*) = 1 - \alpha$ and α is typically 5%.

Monday, 22 October 2012

Lecture 11 - Content

- χ^2 **goodness of fit tests**
- **Further applications of** χ^2 **GoF tests**

χ^2 Goodness of fit tests

Motivational setting

- Suppose we have n independent trials with X successes and $n - X$ failures: $X \sim \mathcal{B}(n, p)$.
- Test $H_0 : p = p_0$ against $H_1 : p \neq p_0$. If H_0 is true then

	Success	Failure	Total
Observed	$O_1 = X$	$O_2 = n - X$	n
Expected	$E_1 = np_0$	$E_2 = n(1 - p_0)$	n

- Large values of $|X - np_0|$ support H_1. Thus large values of

$$\tau = \frac{(X - np_0)^2}{np_0(1 - p_0)}$$

support H_1.

Motivational setting (cont)

- Note,

$$\begin{aligned}(O_2 - E_2)^2 &= [(n - X) - (n - np_0)]^2 \\ &= (X - np_0)^2 = (O_1 - E_1)^2.\end{aligned}$$

- Also,

$$\frac{1}{np_0} + \frac{1}{n(1 - p_0)} = \frac{1}{np_0(1 - p_0)}.$$

- Thus,

$$\begin{aligned}\tau &= \frac{(X - np_0)^2}{np_0(1 - p_0)} = (X - np_0)^2[\frac{1}{np_0} + \frac{1}{n(1 - p_0)}] \\ &= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}.\end{aligned}$$

- This is a special case of Pearson's χ^2 statistic.

Pearson's X^2 GoF test

Assume we have g categories, not just success/failure and H_0 specifies a model giving expected frequencies for each category.

Definition 16. Pearson's χ^2 test-statistic is

$$X^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed frequency in the ith category and E_i is the expected frequency if H_0 is true.

Alternative Calculation Formula for X^2

Pearson's X^2 GoF test (cont)

- Thus, the easiest form for calculation purposes of X^2 is,

$$X^2 = \sum_{i=1}^{g} \frac{O_i^2}{E_i} - n.$$

- We reject the model if the X^2-statistic is too large.
- The sampling distribution of the statistic has (asymptotically) a chi-squared distribution with $g - 1$ degrees of freedom.

$$P\text{-value} = \mathrm{P}(\chi^2_{g-1} \geq \text{observed } X^2 \text{ value}).$$

  - Note that $\chi^2_1 = Z^2$, where $Z \sim \mathcal{N}(0, 1)$. In R with `pchisq()` or tables.
  - The X^2 test should only be used when the expected frequencies, E_i, are greater than 5.
    (Recall this corresponds to $np \geq 5$ for the normal approximation to the binomial!)

The χ^2 distribution

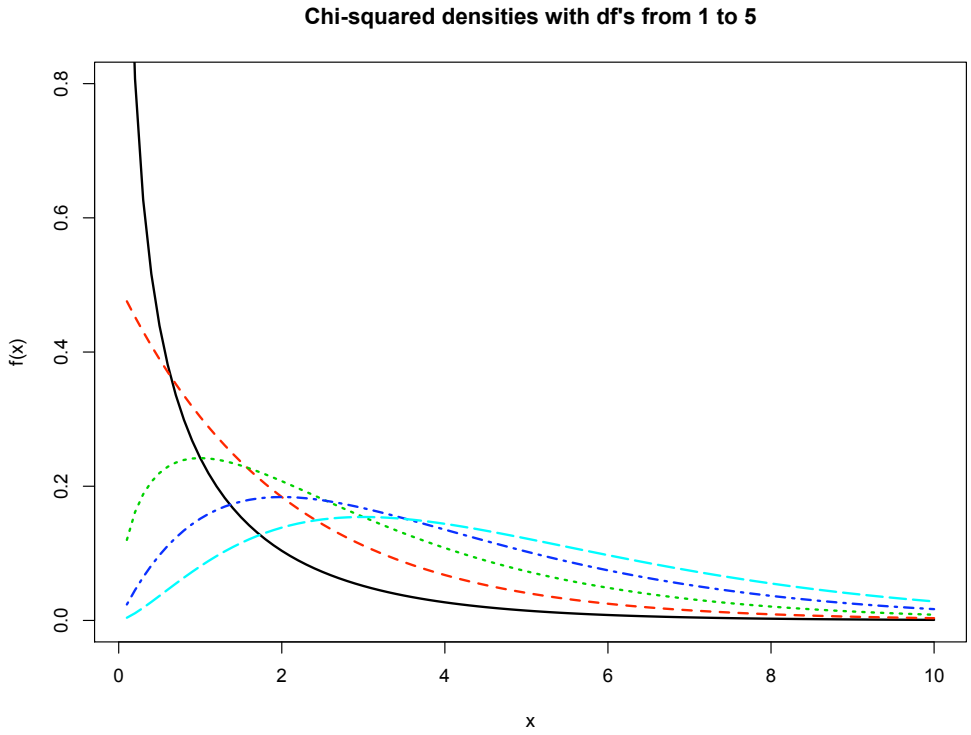
- The χ^2 r.v. can only take non-negative values. The distribution is not symmetric but is right skewed. Tables typically give

$$\mathrm{P}(\chi^2_\nu > x) = p = \texttt{1-pchisq(x,nu)}$$

  for particular d.f. ν and p values.
- With R we can visualise the densities:

```
> x = 1:100/10
> plot(x,dchisq(x,1),type="l",ylim=c(0,0.8))
> points(x,dchisq(x,2),type="l",lty=2,col=2)
> points(x,dchisq(x,3),type="l",lty=3,col=3)
> points(x,dchisq(x,4),type="l",lty=4,col=4)
> points(x,dchisq(x,5),type="l",lty=5,col=5)
```

The χ^2 distribution (cont)



Example.

(i) $\mathrm{P}(\chi^2_1 > 3.841) = ?$

- From tables the points tabulated below are x, where $P(\chi^2_\nu > x) = p$

	p					
ν	0.25	0.15	0.10	0.05	0.025	0.01
1	1.323	2.072	2.706	3.841	5.024	6.635

- Using R:

```
> 1-pchisq(3.841,1)
[1] 0.05001368
> pchisq(3.841,1,lower.tail=FALSE)
[1] 0.05001368
```

(Note also that $1.96^2 = 3.841,\ P(|Z|^2 > 1.96^2) = 0.05$.)

(ii) $\mathrm{P}(\chi^2_{10} > 20) = ?$

- From tables

	p					
ν	0.25	0.15	0.10	0.05	0.025	0.01
10	12.549	14.534	15.987	18.307	20.483	23.209

  So that $0.025 < \mathrm{P}(\chi^2_{10} > 20) < 0.05$.

- Via R

```
> pchisq(20,10,lower.tail=FALSE)
[1] 0.02925269
```

(iii) $\mathrm{P}(13.848 < \chi^2_{24} < 39.364) = ?$

- Can't be obtained because tables only have $1 \leq \nu \leq 10$.
- Via R

```
> pchisq(39.364,24) - pchisq(13.848,24)
[1] 0.9250087
```


Example (Phenotypes, PQ p117)**.** In an experiment involving a dihybrid cross of flies, 148 progeny were classified by phenotype as follows.

AB	Ab	aB	ab	Total
87	31	25	5	148

- Genetic theory predicts a ratio 9:3:3:1 for AB:Ab:aB:ab.
- Do the data support the theory?
-

Example (cont).

-

-

```
pchisq(2.775,3,lower.tail=FALSE)
```

-

Example (Accidents). The number of fatal accidents on NSW roads in the July month between 2004 - 2010 were:

2004	2005	2006	2007	2008	2009	2010
44	50	34	41	34	27	29

Test the claim that the accident rate didn't change over this seven year period:

- p_i denotes the probability that a fatal accident is 'allocated' to month i.
- Model: $p_i = \frac{1}{7}, \quad i = 1, \ldots, 7.$
- The total number of accidents is 259. Thus $E_i = \frac{259}{7} = 37.$

  The test statistic is $X^2 = \sum_{i=1}^{7} \frac{O_i^2}{E_i} - 259 = 11.24.$

- Thus, the P-value $= 0.081 \Rightarrow$ no rejection of the claim that the accident rate is constant across the years based on these data.

Further applications of χ^2 GoF tests

- Recall that for known parameters,

$$X^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{g} \frac{O_i^2}{E_i} - n \overset{\text{under } H_0}{\sim} \chi^2_{g-1}.$$

- If we want to check the fit of a model that involves unknown parameters we first have to estimate the parameters.
- Since we use the same data to estimate the parameters and test the fit we find the sampling distribution of the X^2 statistic has to be adjusted.
- The distribution is still χ^2 but the dfs are reduced to $g - k - 1$, where
  - g is the number of categories and
  - k is the smallest number of parameters that need to be estimated using the data.

Example (More on phenotypes)**.** In a backcross experiment to investigate the genetic linkage between two factors A and B in a species of flower, some researchers classified 400 offspring by phenotype as follows:

AB	Ab	aB	ab
128	86	74	112

(i) Under the no linkage model, the four phenotypes are equally likely.
Show that this model is a poor fit.

(ii) If linkage is in the coupling phase, the probabilities of the four phenotypes are

AB	Ab	aB	ab
$\frac{1}{2}(1-p)$	$\frac{1}{2}p$	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$

where p is the 'recombination fraction' and is estimated by the overall proportion of Ab and aB.
Show that this model fits the data well.

Example (cont; (i)). Model says that all categories are equally likely.

- The total number of observations is $n = 400$.
- Observed O_i : 128 86 74 112
  Expected E_i : 100 100 100 100
- $X^2 = \sum_i \frac{O_i^2}{E_i} - n = 18.$
- $P = \mathrm{P}(\chi^2_3 \geq 18) < 0.01.$
- Thus the data are not consistent with the model.

Example (cont; (ii)). Here, $\widehat{p} = (86 + 74)/400 = 0.4$.

- The expected frequencies are

$$E_1 = E_4 = 400 \times 0.3 = 120$$

$$E_2 = E_3 = 400 \times 0.2 = 80.$$

- $X^2 = 1.97$ and $g - k - 1 = 2$
- $P = \mathrm{P}(\chi^2_2 \geq 1.97) > 0.10.$
- The data are consistent with this model.

Tuesday, 23 October 2012

Lecture 12 - Content

- **Further applications of χ^2 GoF tests (cont)**

Example (Infections)**.** 200 groups of 5 insects each were inspected. For each group the number of infected insects (x) was counted giving:

$$3,\ 2,\ 5,\ 1,\ 0,\ ..,\ 2.$$

The data were condensed into the table below, writing x_i for the number infected and f_i for the corresponding frequency:

x_i :	0	1	2	3	4	5	Total
f_i :	20	62	55	38	20	5	200

Does the binomial model fit the data?

- The null hypothesis is that $X \sim \mathcal{B}(5, p)$. We need to estimate p.
- There were $5 \times 200 = 1000$ insects in total and 391 of these were infected, i.e. an estimate is

$$\widehat{p} = \frac{391}{1000} = 0.391.$$

Example (cont)**.**

i	0	1	2	3	4	5	Total
p_i	0.0837	0.2689	0.3453	0.2217	0.0712	0.0091	1.00
E_i	16.754	53.783	69.061	44.340	14.234	1.828	200
O_i	20	62	55	38	20	5	200

Notice that the E_i value falls below 5 for the last group.

Testing the fit of a normal model

Given a data set $x_1, x_2, .., x_n$ we want to test if the data come from a $\mathcal{N}(\mu, \sigma^2)$ population.

(a) First calculate the sample mean, $\overline{x}$, and the sample variance, s^2.

(b) Form a grouped frequency table summary of the data with (ideally) 5 to 10 categories. Aim to have at least 5 values in each category.

(c) To check the normal claim work out the expected frequencies for each category by fitting $\mathcal{N}(\overline{x}, s^2)$.

(d) Use X^2 as the test statistic. To calculate the P-value use $(g - 2 - 1)$ df.

Example (Rainfall). We have $n = 30$ observations corresponding to Sydney's annual rainfall (in inches) from 1980-2009 (from http://www.bom.gov.au):

```
> y = c( 956,1083,1499, 994, 816, 995,1200, 860,1359, 822,
+       1470,1649,1078,1149,1230, 907, 913,1282,1121,1977,
+       1526,1862,1313,1225,1217,1801,1346, 838,1038, 736)
> x = 2009:1980
```

Test if the rainfall follows a normal distribution.

(a) $\overline{y} = 1208.733$ and $s^2 = 105762.8$.

(b) Grouping the data into a frequency table:

Interval	Frequency
$y \leq 900$	5
$900 < y \leq 1200$	11
$1200 < y \leq 1500$	9
$y > 1500$	5

Example (cont).

(c) We now calculate the expected frequencies using

$$Y \sim \mathcal{N}(1208.733, 325.212^2)$$

$$P(Y \leq 900) = \mathrm{P}\left(Z \leq \frac{900 - 1208.733}{325.212}\right) = \mathrm{P}(Z \leq -0.95) = 0.1712.$$

Thus $E_1 = 30 \times 0.1712 = 5.137$.

$$\begin{aligned} P(900 < Y \leq 1200) &= P(-0.95 < Z \leq -0.03) \\ &= 0.318 \end{aligned}$$

$$E_2 = 30 \times 0.318 = 9.542.$$

Similarly,
$E_3 = 30 \times 0.3254 = 9.765$ and
$E_4 = 30 - 5.137 - 9.542 - 9.765 = 5.556$.

Example (cont). (d) Our table is

Interval	Frequency	Expected
$y \leq 900$	5	5.137
$900 < y \leq 1200$	11	9.542
$1200 < y \leq 1500$	9	9.765
$y > 1500$	5	5.556

The test statistic is

$$X^2 = \frac{5^2}{5.137} + \frac{11^2}{9.542} + \frac{9^2}{9.765} + \frac{5^2}{5.556} - 30 = 0.342.$$

Here $g = 4$ and $k = 2$ so we have 1 d.f.

The P-value is $P(\chi^2_1 \geq 0.342) = 0.559,$ with R.

Thus the data are consistent with the normal model.

Tests for independence

If we have data classified according to two attributes then we can construct a contingency table which is just a convenient way of presenting the group frequencies. For example, we have data on 422 drivers and motorcyclists killed in NSW in 1988. We classify the people by blood alcohol level and gender.

Alc (g/100ml)	0	$(0, 0.08)$	$[0.08, 0.15)$	≥ 0.15	Total
Male	206	37	35	76	354
Female	53	5	4	6	68
Total	259	42	39	82	

Test the claim that gender is independent of blood alcohol level.

A probability model for contingency tables

- Let p_{ij} denote the probability of a victim being gender i and alcohol level group j then the independence model says:

$$p_{ij} = p_i^g p_j^a, \quad \text{where}$$

  - p_i^g is the prob. of being of gender i,
  - p_j^a is the prob. of being in alcohol group j.

- We estimate p_i^s and p_j^a by the marginal proportions,

$$\widehat{p}_1^g = \frac{354}{422} \Rightarrow \widehat{p}_2^g = 1 - \widehat{p}_1^g = \frac{68}{422};$$

$$\widehat{p}_1^a = \frac{259}{422}, \ \widehat{p}_2^a = \frac{42}{422}, \ \widehat{p}_3^a = \frac{39}{422}, \ \text{and} \ \widehat{p}_4^a = \frac{82}{422}.$$

- The expected frequency under the independence model in the Male/Alcohol 0 group is

$$422 \times \widehat{p}_{11} = 422 \times \frac{259}{422} \times \frac{354}{422} = 217.265.$$

- For a general table with entries x_{ij} the expected frequencies under independence are

$$E_{ij} = n \times \frac{x_{i\bullet}}{n} \times \frac{x_{\bullet j}}{n} = \frac{x_{i\bullet} \times x_{\bullet j}}{n},$$

where $x_{i\bullet}$ denotes the sum of the ith row and $x_{\bullet j}$ denotes the sum of the jth column.

- The expected frequencies for the accident data are

```
217.265   35.232   32.716   68.787
 41.735    6.768    6.284   13.213
```

-

-

```
1-pchisq(9.859,3)
```

-

- In general if we have a table with r rows and c columns then the test statistic for testing independence will have

$$(r-1)(c-1) \text{ degrees of freedom.}$$

Tests for Symmetry in Tables

The following data for 205 married people were reported in Yule (1900).

	Wife		
Husband	Tall	Medium	Short
Tall	18	28	14
Medium	20	51	28
Short	12	25	9

Here there are $g = 9$ groups. Suppose we wish to test for table symmetry, i.e. the probability of Tall Men marrying Short Women and Short Men marrying Tall Women are roughly equal.

The table corresponding to a symmetric table model would have $k = 5$ parameters

	Wife		
Husband	Tall	Medium	Short
Tall	p_1	p_2	p_3
Medium	p_2	p_4	p_5
Short	p_3	p_5	$1 - p_1 - \cdots - p_5$

Tests for Symmetry in Tables

	Wife		
Husband	Tall	Medium	Short
Tall	18	28	14
Medium	20	51	28
Short	12	25	9

Under the symmetric model the expected values of the table entries would be

	Wife		
Husband	Tall	Medium	Short
Tall	18	24	13
Medium	24	51	26.5
Short	13	26.5	9

Tests for Symmetry in Tables

The value of the test statistic would be

$$X^2 = \frac{(18-18)^2}{18} + \frac{(28-24)^2}{24} + \cdots + \frac{(9-9)^2}{9} = 1.656991$$

The degrees of freedom would be $\nu = g - k - 1 = 9 - 5 - 1 = 3$.

The P-value would then be

$$P(\chi^2_3 > 1.656991) = \texttt{1-pchisq(1.656991,3)} = 0.6465376$$

So we would accept the null hypothesis for table symmetry.

Note that for a test for independence the test statistic is 2.907 (check yourself!) with $\nu = 4$ degrees of freedom and a P-value of 0.5735. Thus, a null hypothesis of independent rows and columns is also consistent with the data!

THE END !!!