Semester 2, 2012 (Last adjustments: October 17, 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

Monday, 17 September 2012

# Lecture 1 - 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

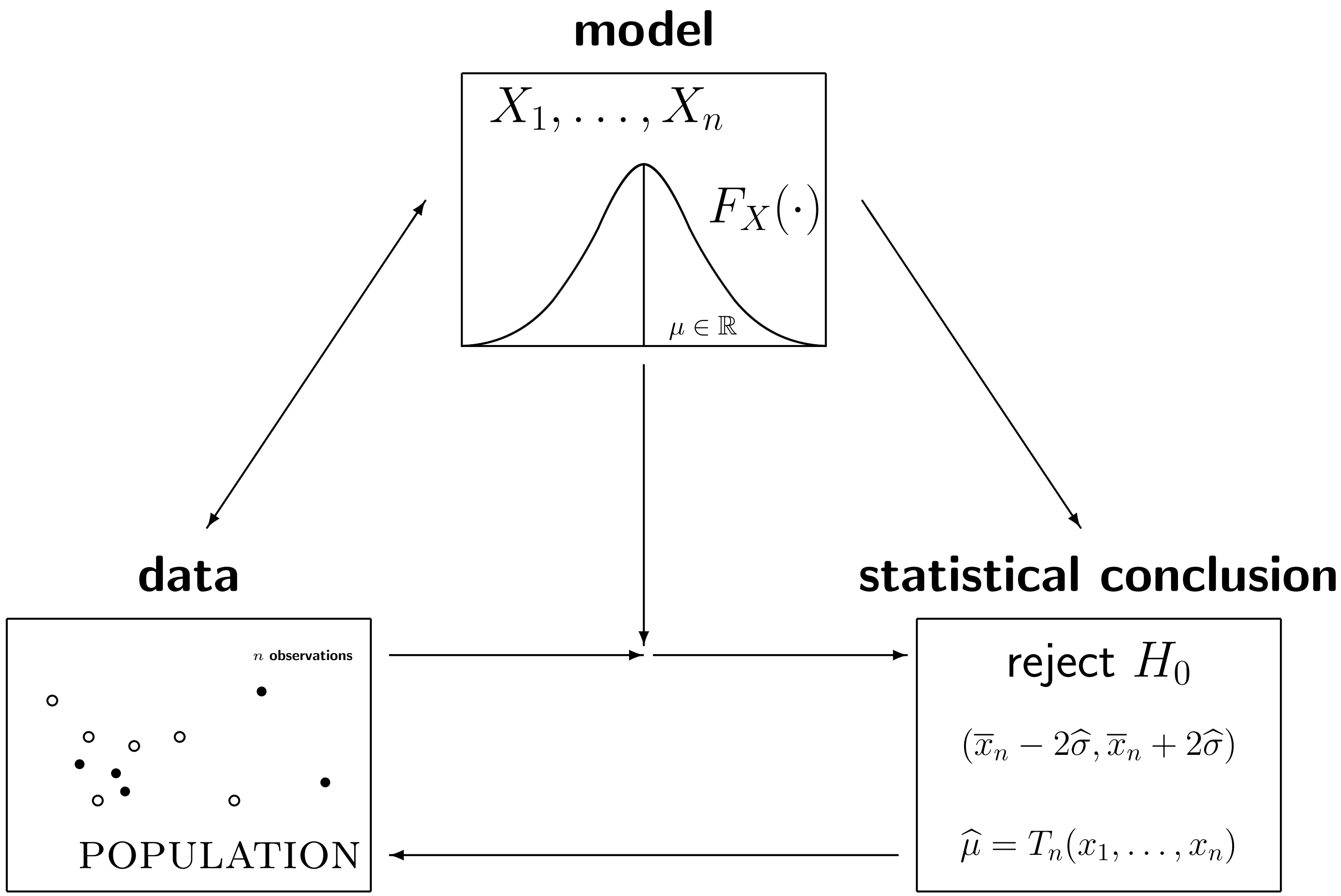
Statistical inference is inference about a population from a random sample drawn from it.

**Definition 1.** 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 2.** A hypothesis, $H$, is a statement about an unknown parameter (e.g. $\mu$) 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 3.** 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 4.** 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, $\tau$ (some function of our observed data).
- Find the distribution of $\tau$ 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 5.** 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

$$H_1\colon p < 0.5$$

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

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?

Let $X$ denote the number not getting the flu in the second year.

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

- Null hypothesis: $H_0 : p = 0.25$.
- Alternative hypothesis: $H_1 : p > 0.25$.
- Is the above observation unusual if $H_0$ is true?
- Large values of $X$ support $H_1$. We observe 9 not getting the flu.

- We can approximate $X$ by the normal $Y \sim \mathcal{N}(5, 15/4)$ if $H_0$ is true.

$$\begin{aligned} \mathrm{P}(X \geq 9) &= 1 - \mathrm{P}(X \leq 8) \\ &\simeq 1 - \mathrm{P}\left(Z \leq \frac{8.5 - 5}{\sqrt{15/4}}\right) \\ &= 1 - \Phi(1.807) \\ &= 1 - 0.9649 = 0.0351. \end{aligned}$$

(The exact value is 0.041.)

- Thus if $H_0$ is true then we have observed a 'rare' event.

## 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. Why? When?
- 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, $H_0$. The claim against which evidence is searched for.
   - ☐ Alternative hypothesis, $H_1$. The alternative you will consider if $H_0$ is false.
2. What is the test statistic, $\tau$, and its sampling distribution if $H_0$ is true.
3. What is the critical region of the test statistic, i.e. which values of $\tau$ argue against $H_0$?
4. Observed test statistic (value of $\tau$ 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 2 - Content

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

## Checklist for statistical tests

1. Hypotheses:
   - □ Null hypothesis, $H_0$. The claim against which evidence is searched for.
   - □ Alternative hypothesis, $H_1$. The alternative you will consider if $H_0$ is false.
2. What is the test statistic, $\tau$, and its sampling distribution if $H_0$ is true.
3. What is the critical region of the test statistic, i.e. which values of $\tau$ argue against $H_0$?
4. Observed test statistic (value of $\tau$ 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

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

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

$$H_1\colon p \neq p_0$$

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

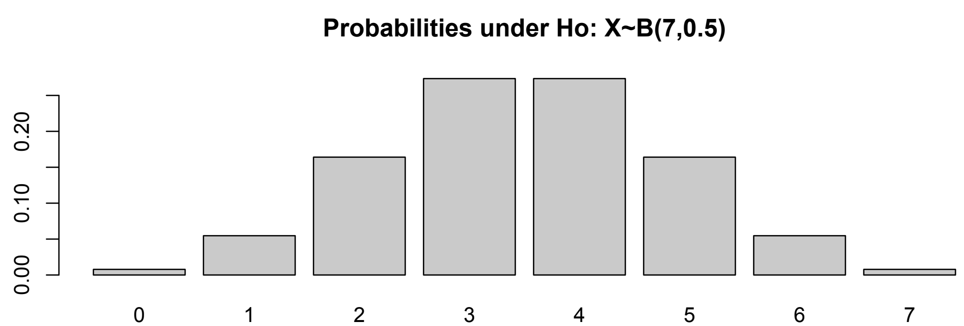
$$H_0 \colon p = p_0$$

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 $H_0$:

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



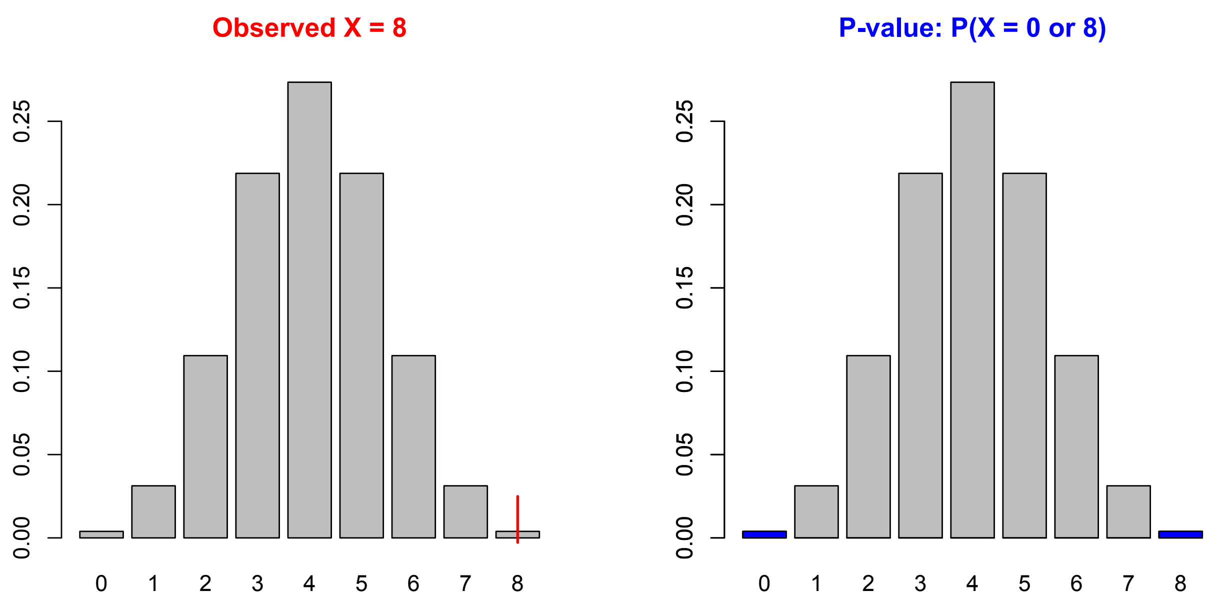
**Example** (Paul the octopus)**.** Is Paul the octopus guessing?

(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 $= 8$'.
- 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).




- $\mathrm{P}(X \le 0) + \mathrm{P}(X \ge 8) =$ `2*pbinom(0,8,0.5)` $=$ 0.0078125.
- Conclusion: 8 correct predictions out of 8 attempts does provide sufficient evidence to make us reject the claim that $p = 0.5$.
- 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?

- Let $X$ = 'number of defectives in the sample of size n = 100'.

$$X \sim \mathcal{B}(100, p) \simeq Y \sim \mathcal{N}(np, np(1-p)) \text{ if } np \geq 5 \text{ and } n(1-p) \geq 5.$$

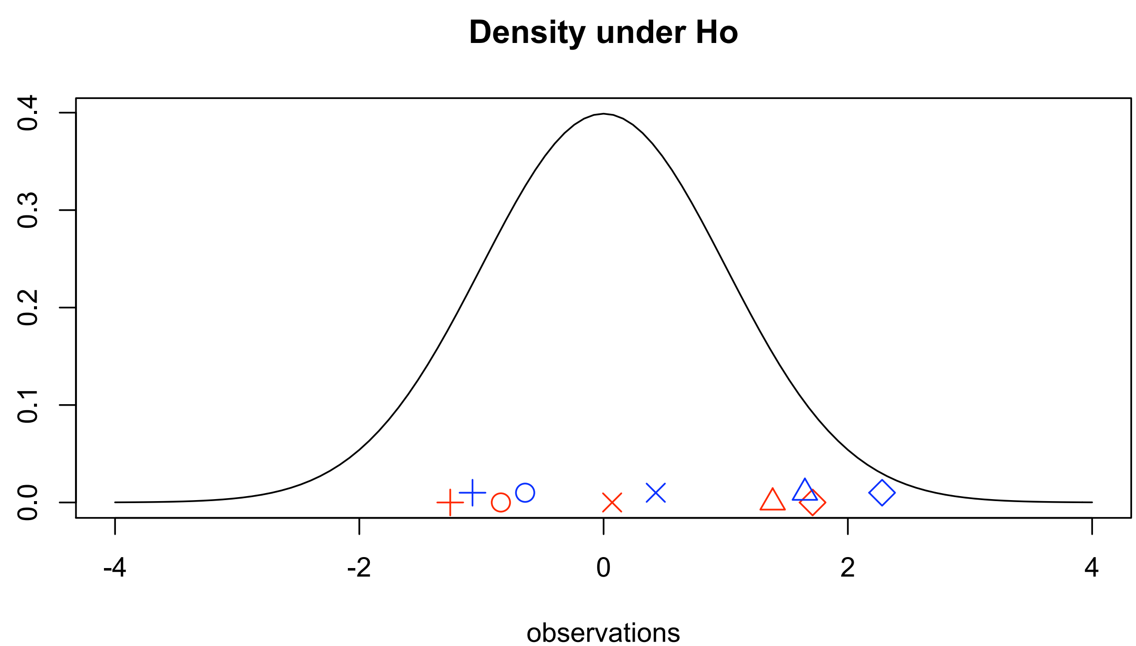
- Test: $H_0 : p = 0.07$ against $H_1 : p \neq 0.07$
- Under $H_0$: $X \sim \mathcal{B}(100, 0.07) \simeq \mathcal{N}(7, 6.51)$ because $\mathrm{E}(X) = 7$.
- $P$-value: $\mathrm{P}(|X-7| \geq 4) \simeq \mathrm{P}(|Z| \geq \underbrace{3.5}_{\text{c.c.}}/\sqrt{6.51}) = 2(1 - \Phi(1.37))$

```
> 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 1.** 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.* Suppose that the probability of $X = x$ and $Y = y$ are defined by $P(X = x, Y = y)$. Due to independence

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Since $X$ and $Y$ are identically distributed

$$P(X = x, Y = y) = P(X = x)P(Y = y) = P(Y = x)P(X = y).$$

Using independence (in reverse)

$$P(X = x, Y = y) = P(X = x)P(Y = y) = P(Y = x)P(X = y) = P(Y = x, X = y).$$

Then

$$
\begin{aligned}
\mathrm{P}(D=d) &= \sum_x \sum_y \mathbf{1}(x-y=d)\,\mathrm{P}(X=x,Y=y) \\
&= \sum_x \sum_y \mathbf{1}(x-y=d)\,\mathrm{P}(X=y,Y=x) \\
&\quad \text{(using } \mathrm{P}(X=x,Y=y)=\mathrm{P}(X=y,Y=x)\text{)} \\
&= \sum_y \sum_x \mathbf{1}(y-x=d)\,\mathrm{P}(X=x,Y=y) \\
&\quad \text{(switching labels } x \text{ and } y\text{, OK since } X \text{ and } Y \text{ iid)} \\
&= \sum_x \sum_y \mathbf{1}(x-y=-d)\,\mathrm{P}(X=x,Y=y) \\
&\quad \text{(reorder sum and multiply condition by -1 in } \mathbf{1}\text{)} \\
&= \mathrm{P}(D=-d).
\end{aligned}
$$

Hence, $P(D=X-Y=-d)=P(D=X-Y=d)$, i.e. the distribution of $D$ is symmetric. $\square$

## 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).

- The 10 differences are: $0.4, -0.2,\ 0.5,\ 0.7, -0.1,\ 0.5,\ 0.6,\ 0,\ 0.2, -0.1$
- Base the test on $X$, the number of positive differences in the $m = 9$ non-zero differences.
- Note we ignore differences that are 0!
- Let $p_+$ be the probability of a positive difference.
- Express the hypotheses in terms of $p_+$.
- $H_0 : p_+ = \frac{1}{2}$ against $H_1 : p_+ > \frac{1}{2}$.

$$X \sim \mathcal{B}(9, 0.5).$$

- There are 6 positive differences in the sample,

$$P\text{-value} = \mathrm{P}(X \geq 6) = 1 - \mathrm{P}(X \leq 5) = 0.2539.$$

- Thus, based on the sign test, the data are consistent with $H_0$.

## **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 $\mu$.

Monday, 1 October 2012

## Lecture 3 - Content

- **No lecture due to Labour Day holiday**

Tuesday, 2 October 2012

## Lecture 4 - Content

- **Tests for the mean $\mu$**
- **$Z$-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 $\alpha$ (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 $\alpha$ 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 $\mu$

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 $\mu$

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 $\mu$ is $\tau = \overline{X}$, the sample mean.

### Two scenarios

At this point it is important to distinguish between two situations

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

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

## Assumption 3: $\sigma$ is known

The $Z$-test is constructed under the assumption that $\sigma$ is known.

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

If $\sigma$ 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 $\mu_0$ 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

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

**Definition 6.** 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:

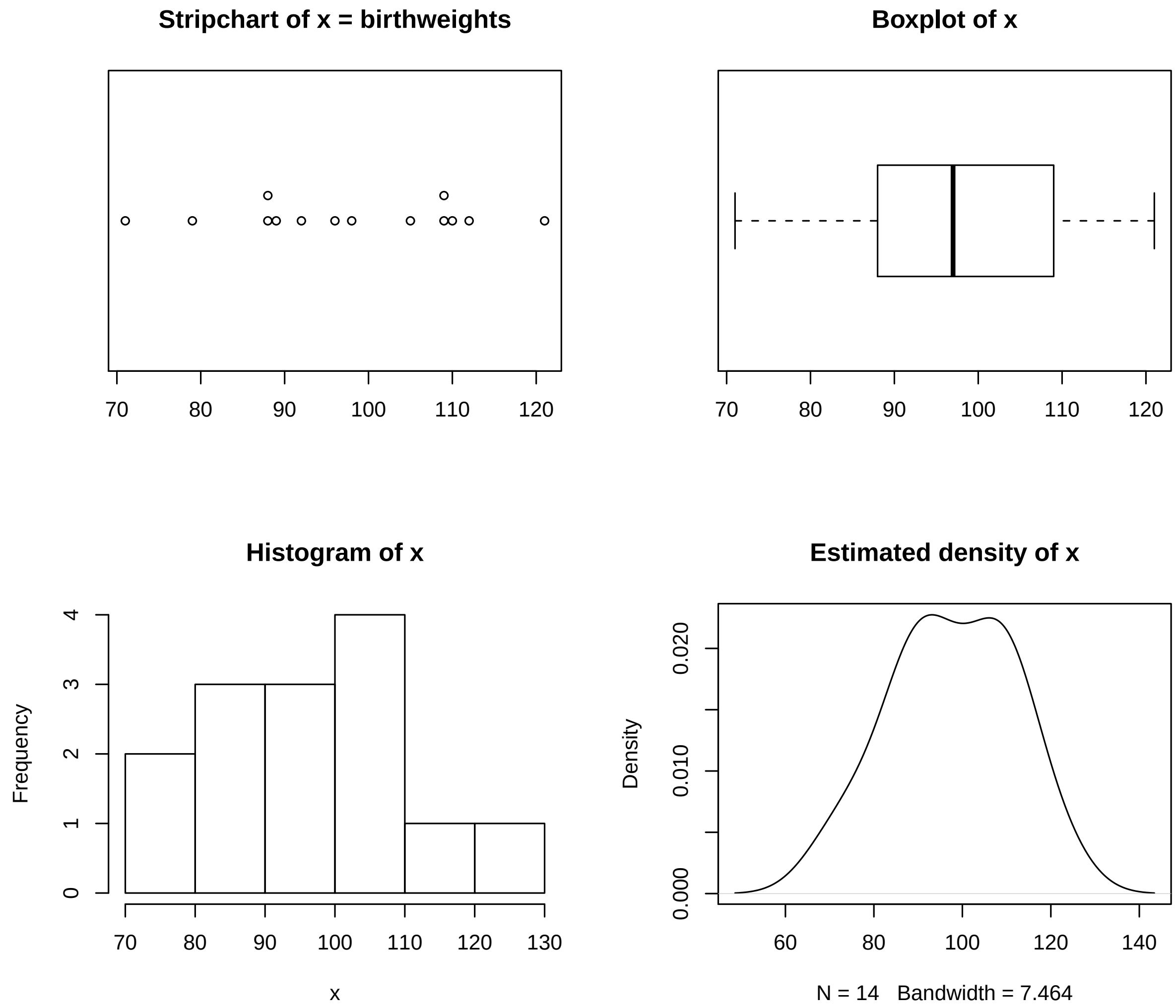
$$79,\ 92,\ 88,\ 98,\ 109,\ 109,\ 112,$$
$$88,\ 105,\ 89,\ 121,\ 71,\ 110,\ 96.$$

- ☐ 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)

- Thus, the $P$-value is

$$\begin{aligned}
P\text{-value} &= \mathrm{P}\left(\overline{W} \leq 97.643\right) \\
&= \mathrm{P}\left(Z \leq \frac{97.643 - 109}{15/\sqrt{14}}\right) \\
&= \mathrm{P}(Z \leq -2.83) \\
&= 1 - \mathsf{pnorm}(97.643, mean = 109, sd = 15/\sqrt{14}) \\
&= 1 - 0.9977 \\
&= 0.0023.
\end{aligned}$$

- Thus there is 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

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

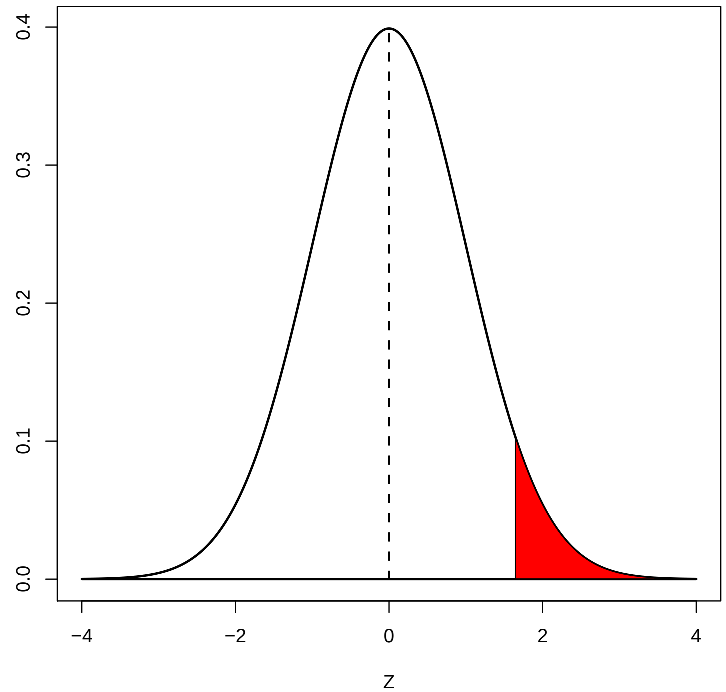
- For normal children of the same age $\mu = 6.04$.
- Is there evidence that the muscle thickness is greater in SIDS victims?

## 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 $\sigma$ is known.

## Two-sided $Z$-tests

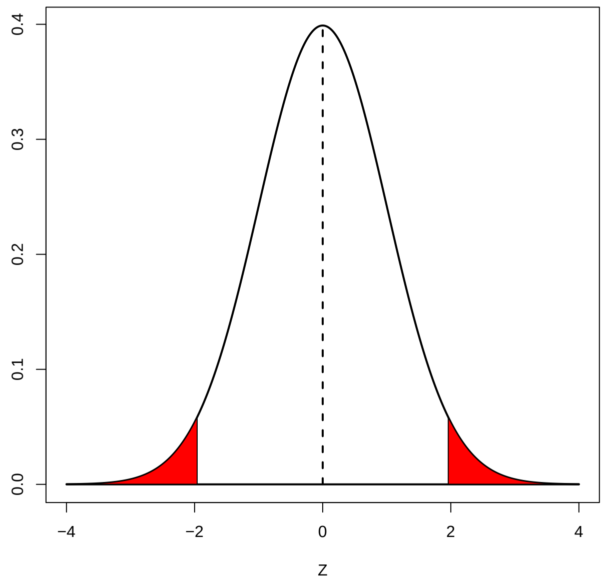
**Example (Breaking strengths).** A new synthetic fishing line is marketed with a manufacturer's claim that the mean breaking strength is 8 kgs with an s.d. of 0.5 kgs. Test this claim if a random sample of 50 lines is tested and the average of the sample of breaking strengths is $\overline{x} = 7.85$ 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 $H_1\colon \mu \neq 8$.
- Large values of $|\overline{X} - 8|$ argue for $H_1$.
- If $H_0$ is true then

$$\overline{X} \sim \mathcal{N}(8, 0.5^2/50).$$

□ The $P$-value is,

$$
\begin{aligned}
P\text{-value} &= \mathrm{P}(|\overline{X} - 8| \geq |7.85 - 8|) = \mathrm{P}\left(|Z| \geq \frac{0.15}{0.5/\sqrt{50}}\right) \\
&= \mathrm{P}(|Z| \geq 2.12) = 2(1 - \Phi(2.12)) = 2(1 - 0.9830) = 0.034.
\end{aligned}
$$



- Thus there is (strong) statistical evidence against $H_0$.

## Conclusions from the previous three examples

- In all of the above examples we have been given the value for the population standard deviation, $\sigma$.
- In practice $\sigma$ 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 $\sigma$ 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 5 - 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 $\sigma$,
- In practice $\sigma$ is generally unknown!
- Estimate $\sigma^2$ 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 2.** If $\overline{X}$ is the mean of a sample of size $n$ taken from a normal distribution having the mean $\mu$ and the variance $\sigma^2$, 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.

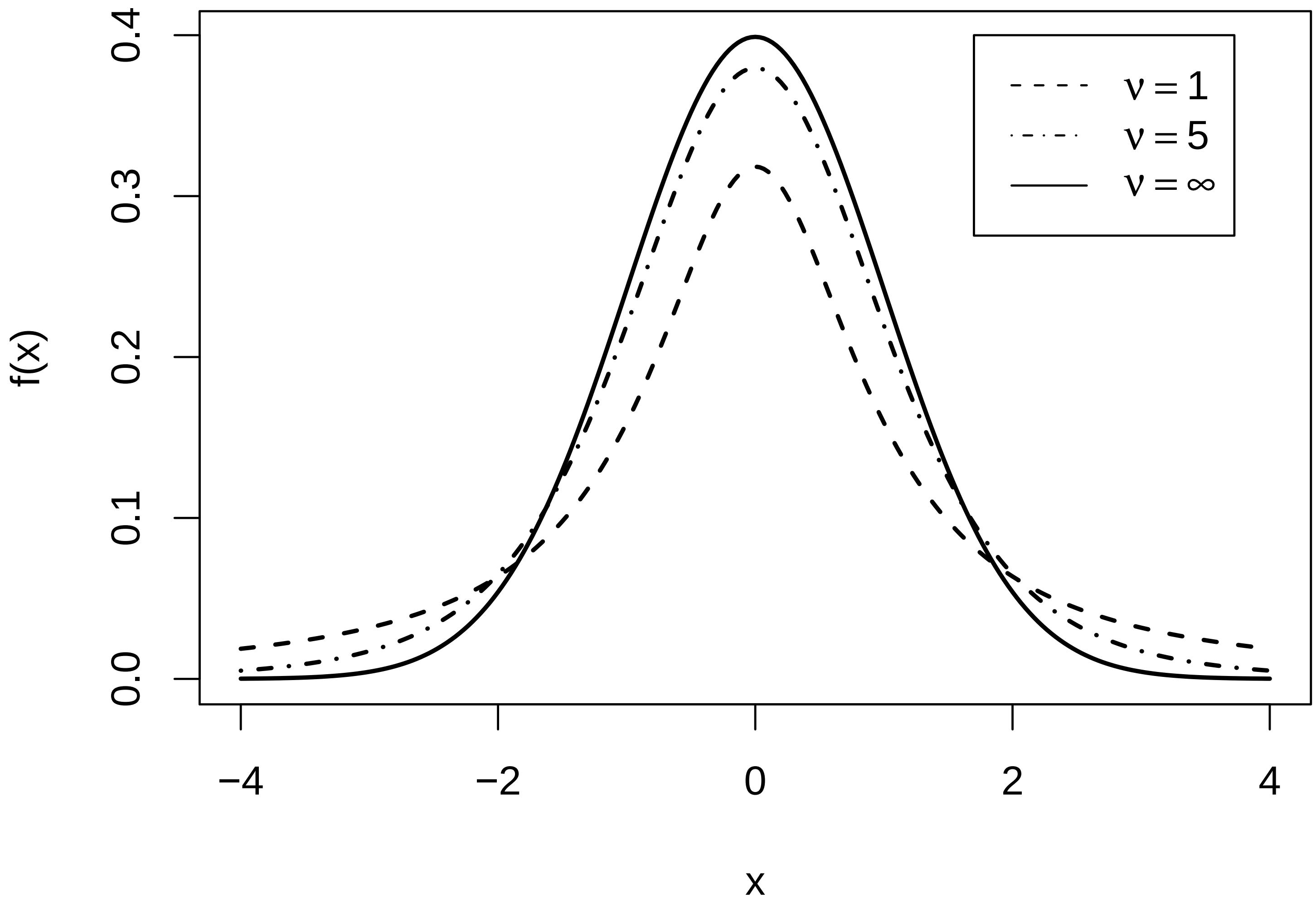
(Note that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.)

## 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 $\sigma$ by $s$.

## The $t$ distribution



## The pdf of the $t$-distribution

**Definition 7.** 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 $\Gamma$-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 $\sigma^2$ is known.

## Assumption 3: population is normal but $\sigma^2$ is unknown

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

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

  where $\nu$ 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.** Find $c$ such that

(i) $\mathrm{P}(t_5 > c) = 0.025$

(ii) $\mathrm{P}(|t_6| \leq c) = 0.90$

(iii) $\mathrm{P}(|t_{27}| \leq c) = 0.95$

(iv) Give bounds for the probability $\mathrm{P}(t_{10} > 2.5)$.

From the table we find

$$\mathrm{P}(t_{10} > 2.228) = 0.025$$
$$\mathrm{P}(t_{10} > 2.764) = 0.01$$

Hence,

$$0.01 < \mathrm{P}(t_{10} > 2.5) < 0.025.$$

**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 $\tau$ support $H_1$.

## Example (cont)

- The observed test statistic is

$$\frac{97.643 - 109}{14.058/\sqrt{14}} = -3.0288$$

- Hence, from tables

$$\mathrm{P}(t_{13} > 3.012) = 0.005$$

and

$$\mathrm{P}(t_{13} > 3.852) = 0.001.$$

- Thus,

$$0.001 < P\text{-value} < 0.005$$

and again we have strong evidence against the null hypothesis $H_0$.

# 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:

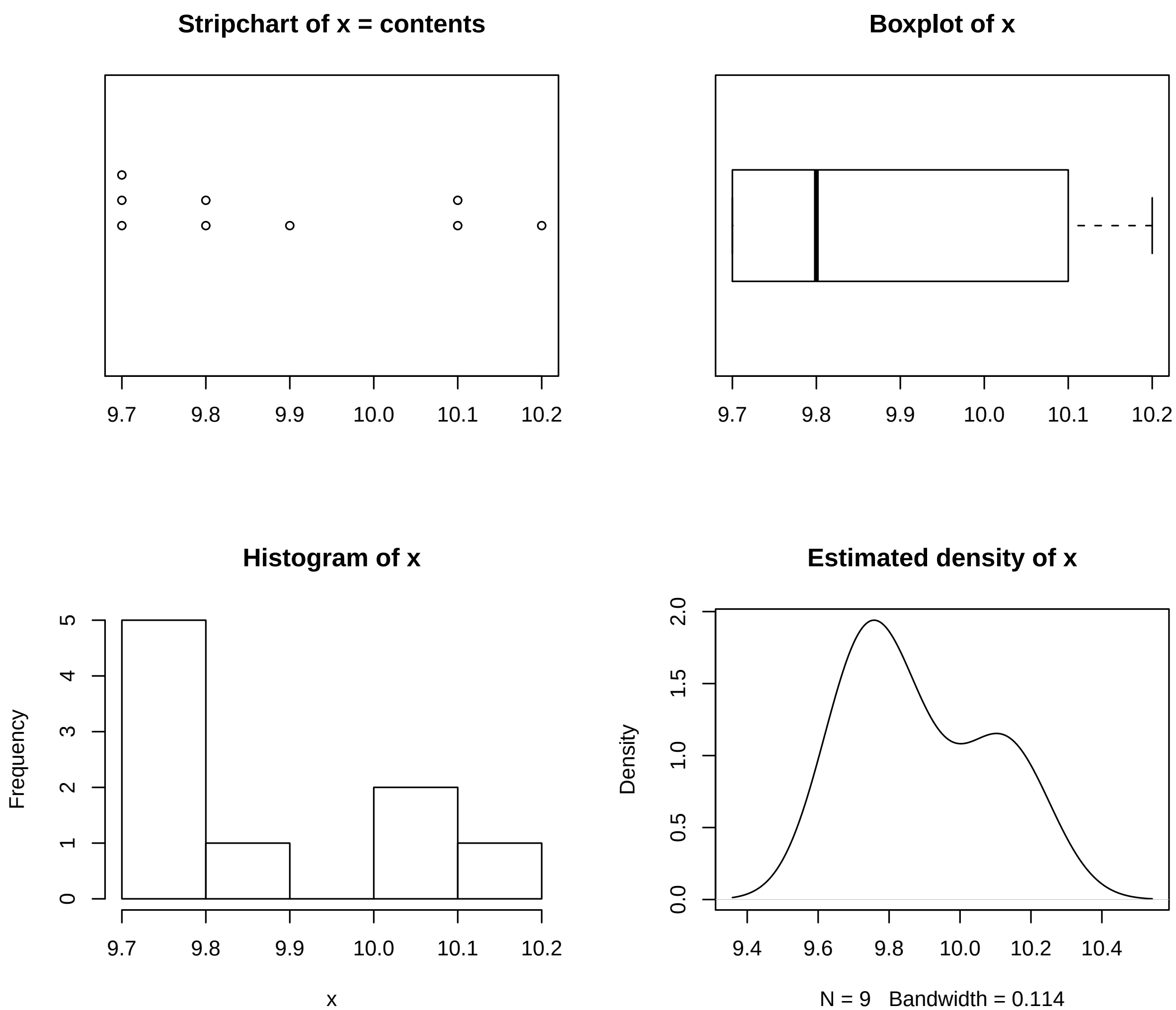
$$10.2,\ 9.7,\ 10.1,\ 9.7,\ 10.1,\ 9.8,\ 9.9,\ 9.8,\ 9.7.$$

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
```

- Test, $H_0 : \mu = 10$ against $H_1 : \mu < 10$.
- Because sample size is very small, base the test on $\tau = \frac{\overline{X}-10}{S/\sqrt{9}}$.
- The observed test statistic value is, $\frac{9.8889-10}{0.1965/3} = -1.696$.
- Small values of $\tau$ support $H_1$: $P$-value $= \mathrm{P}(t_8 < -1.7) = \mathrm{P}(t_8 > 1.7)$.
- From the tables: $\mathrm{P}(t_8 > 1.397) = 0.10$ and $\mathrm{P}(t_8 > 1.860) = 0.05$ so $0.05 < P < 0.10$. (The exact $P$-value is 0.064.)
- Thus there is not strong evidence against the claim that the mean content is 10 litres.

Tuesday, 9 October 2012

# Lecture 6 - Content

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

## References from Phipps & Quine

- Section 3 pages 96–100.

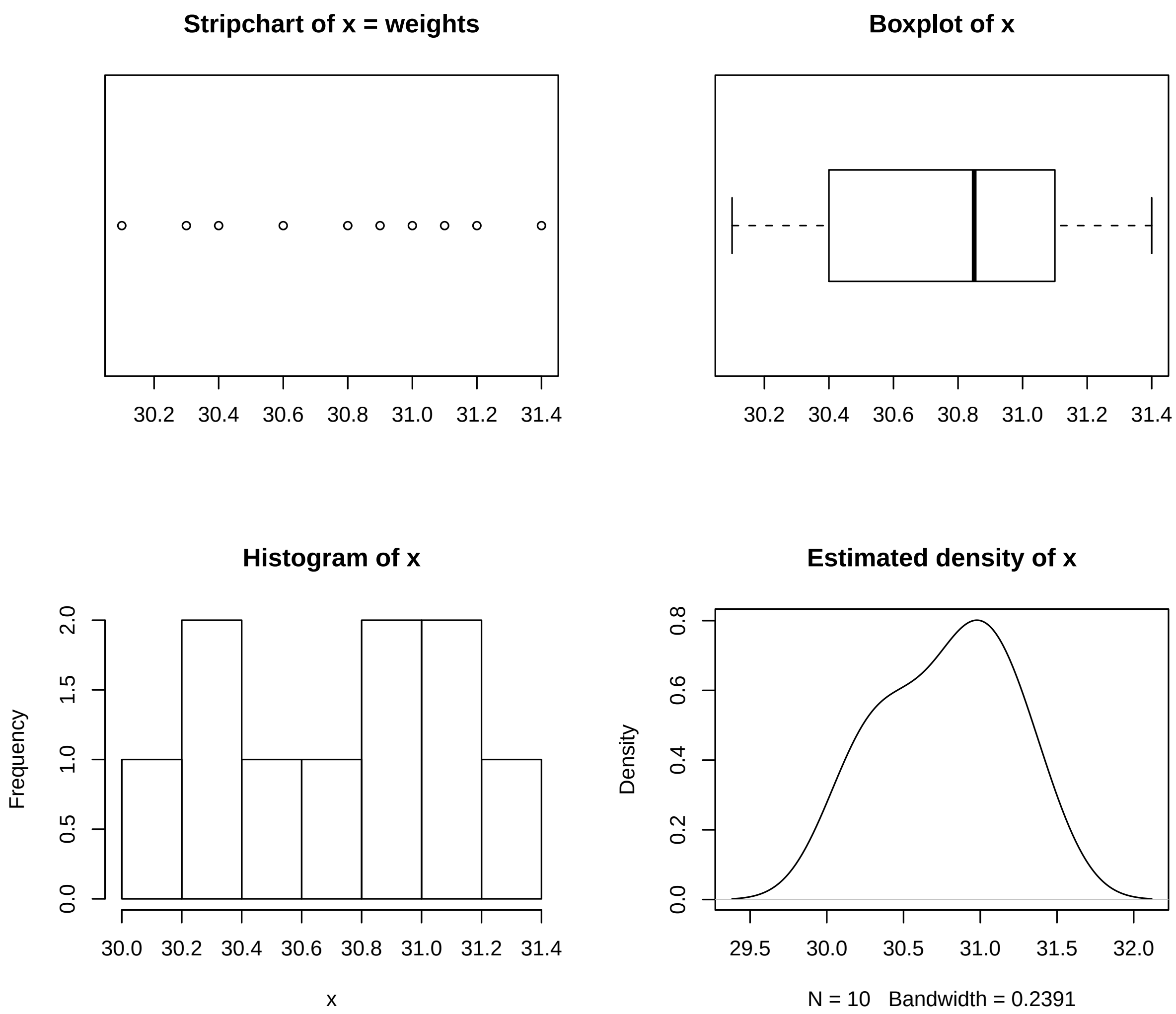
**Example** (Tablets). Ten tablets are weighed giving the weights (in mgs):

```
> 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 $\tau$ 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 $t_{n-1} = t_9$.
- The observed test statistic value is,
$$\frac{30.78 - 30}{0.4211096/\sqrt{10}} = 5.857327.$$
- Because we are performing a 2-sided test, as opposed to a 1-sided test, the $P$-value is
$$2P(t_9 < -5.857327) = 2\,\mathrm{P}(t_9 > 5.857327).$$
- From the tables:
$$\mathrm{P}(t_9 > 4.297) = 0.001.$$

Hence, $\mathrm{P}(t_9 > 5.857327) < 0.001$ and the $P$-value is less than $0.002$ since
$$2\,\mathrm{P}(t_9 > 5.857327) < 0.002.$$

## 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 $0.000241544$.

- Since, the $P$-value is 0.000241544 which is much less than $0.05$ we have strong evidence against $H_0$.
- Hence, there is strong evidence to suggest that the machine producing the tables is set to give a mean weight of 30 mg.

## 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 $\mu_0$ the null hypothesis is $H_0\colon \mu = \mu_0$.
- If $\sigma$ is unknown then $\tau = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$ and

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

- If $\sigma$ is known then $\tau = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ or if $\sigma$ 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

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

## 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 | $X_1$ | $X_2$ | $\dots$ | $X_n$ |
| Y | $Y_1$ | $Y_2$ | $\dots$ | $Y_n$ |
| D | $D_1$ | $D_2$ | $\dots$ | $D_n$ |

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 $X$s or $Y$s except that the $X$s and $Y$s 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:

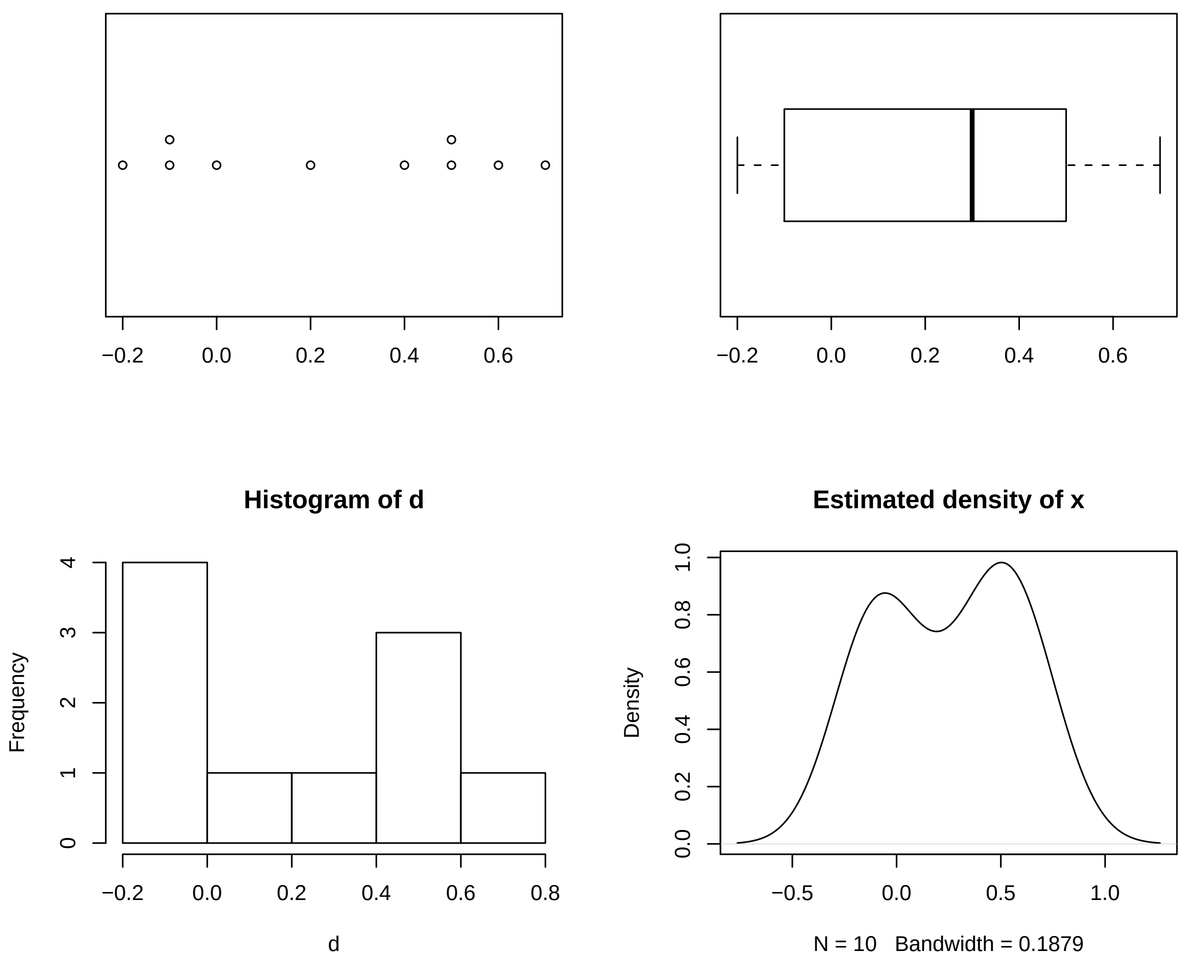
| Rat | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Bioch. | 1.7 | 2.0 | 1.7 | 1.5 | 1.6 | 2.4 | 2.3 | 2.4 | 2.4 | 2.6 |
| Placebo | 2.1 | 1.8 | 2.2 | 2.2 | 1.5 | 2.9 | 2.9 | 2.4 | 2.6 | 2.5 |

- 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 $\tau$ 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 $t_{n-1} = t_9$.
- The observed test statistic value is,

$$\frac{0.25}{0.3308239/\sqrt{10}} = 2.389699.$$

- Because we are performing a 1-sided test the $P$-value is $\mathrm{P}(t_9 > 2.389699)$.
- From the tables:

$$\mathrm{P}(t_9 > 2.262) = 0.025 \quad \text{and} \quad \mathrm{P}(t_9 > 2.821) = 0.01$$

Hence, the $P$-value between $0.025$ and $0.01$.

## 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 $0.02028870$.

- Since, the $P$-value is 0.02028870 which is less than $0.05$ we have evidence against $H_0$.

- Hence, there is evidence to suggest that this experiment does provide evidence of a significant inhibitive effect.

## 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)

- Test $H_0 : \mu = 0$ against $H_1 : \mu > 0$. Here, $\overline{d} = 0.28$ and $s^2 = 0.09956$.
- Test is based on
$$\tau = \frac{\overline{D}}{S/\sqrt{10}}.$$
- Large value of $\tau$ support $H_1$. The observed test statistic is,
$$\frac{0.28}{\sqrt{0.09956/10}} = 2.8062 \Rightarrow P\text{-value} = \mathrm{P}(t_9 \geq 2.8) = 0.001.$$

Monday, 15 October 2012

# Lecture 7 - 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

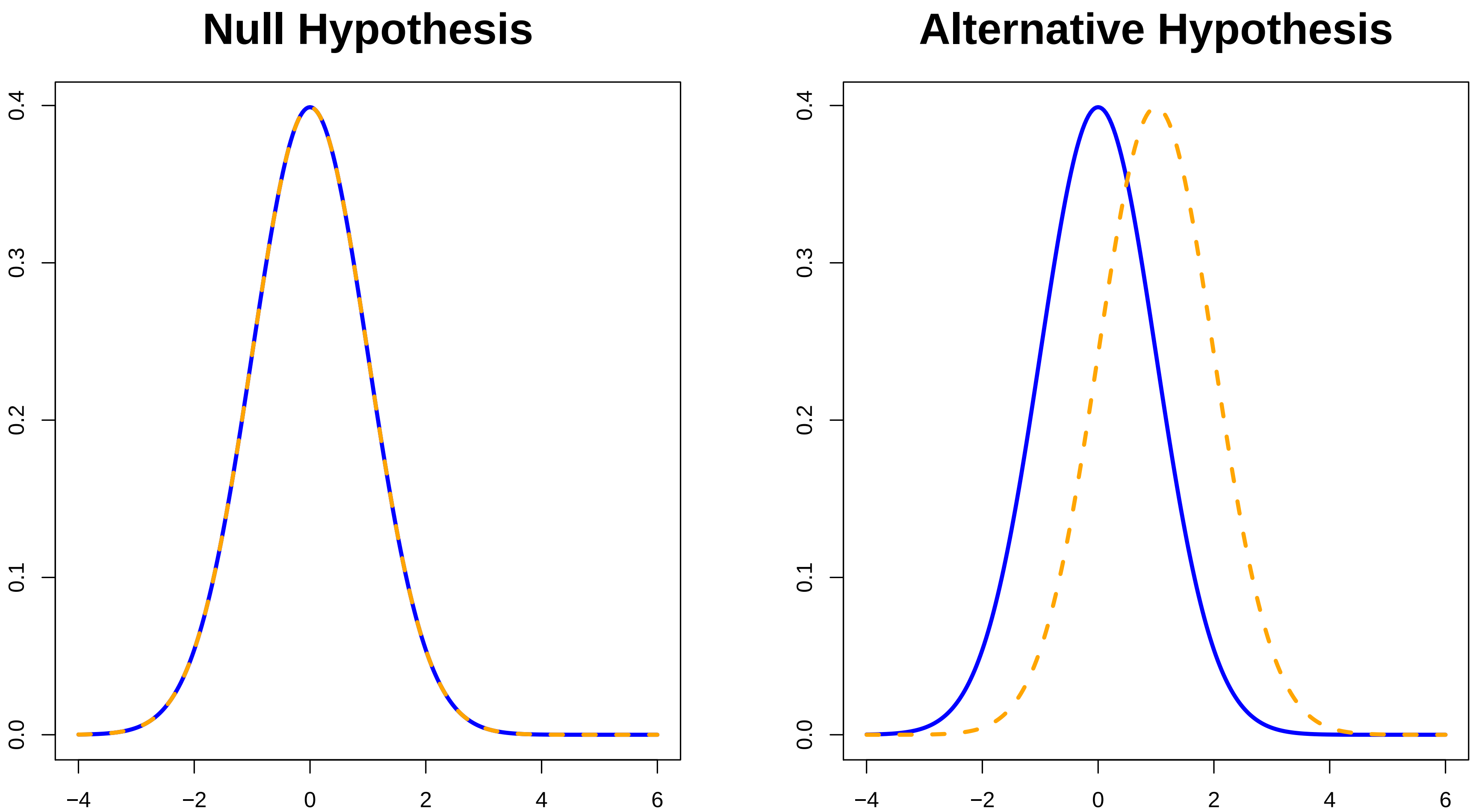
(iii) common variance.

**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 $\sigma^2$ is known then the differences of the means has distribution

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

## Two-sample $Z$- and $t$-test statistics

Hence, if $H_0 : \mu_x = \mu_y$ is true and $\sigma^2$ 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 $\sigma^2$ 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 are taller than women in the MATH1905 tutorial).

**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.

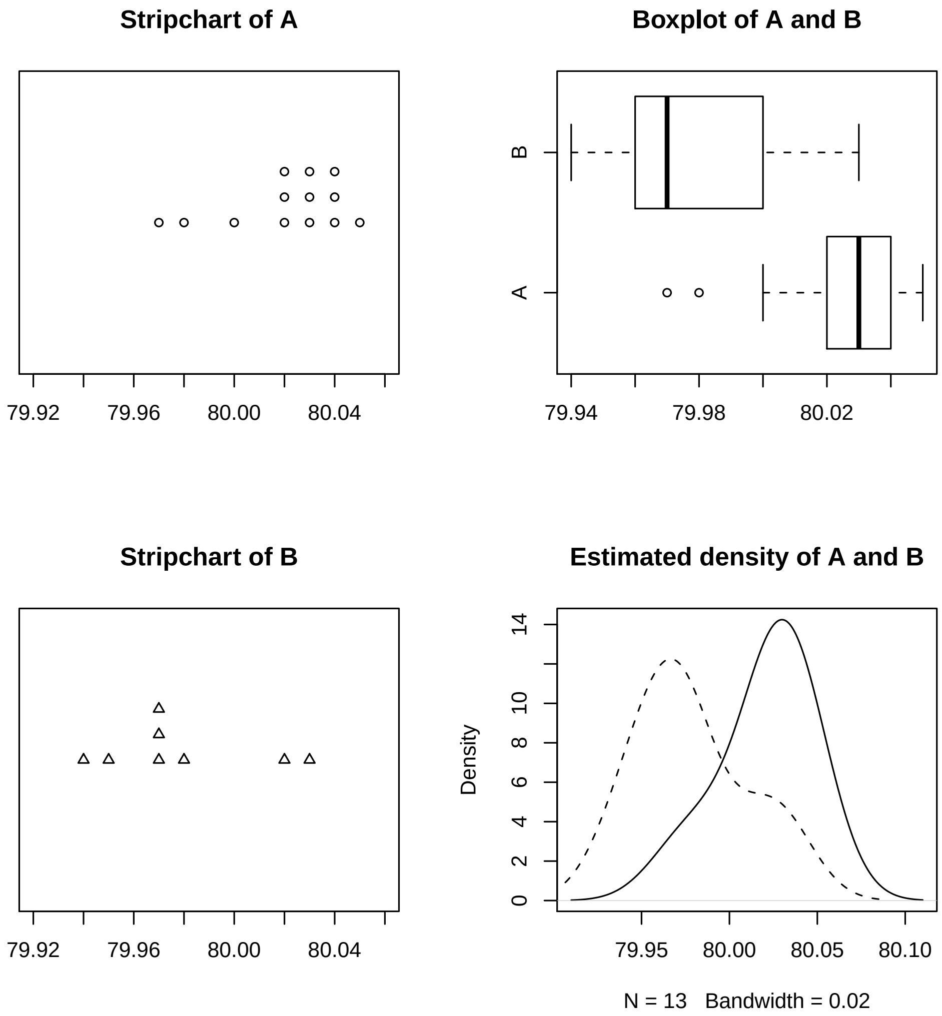
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| A: | 79.98 | 80.04 | 80.02 | 80.04 | 80.03 | 80.03 | 80.04 |
| | 80.05 | 80.03 | 80.02 | 80.00 | 80.02 | 79.97 | |
| B: | 80.02 | 79.94 | 79.98 | 79.97 | | | |
| | 79.97 | 80.03 | 79.95 | 79.97 | | | |

- □ 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: $n_x = 13$ $\overline{x} = 80.0208$ $s_x = 0.02397$

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

- $s_p^2 = \dfrac{12 \times 0.02397^2 + 7 \times 0.03137^2}{19} = 0.02693^2.$
- Test $H_0 : \mu_A = \mu_B$ against $H_1 : \mu_A \neq \mu_B$.
- Large values of $\dfrac{|\overline{X} - \overline{Y}|}{S_p\sqrt{\frac{1}{13} + \frac{1}{8}}}$ support $H_1$
- The observed statistic is $\dfrac{|80.0208 - 79.9788|}{0.02693\sqrt{\frac{1}{13} + \frac{1}{8}}} = 3.47.$
- $P$-value $= \mathrm{P}(|t_{19}| \geq 3.47) = 2\,\mathrm{P}(t_{19} \geq 3.47) \in (0.002, 0.01)$.
- Thus there is strong evidence that the two methods differ.

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 $\sigma$ *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 $\sigma$ *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 $\sigma$ *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 $\mu$

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 $\mu$?
- 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 $\mu$ but with probability one wrong, i.e.

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

- Known result: If $\sigma$ 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 $\mu$.

## 95% CI for $\mu$ if $\sigma$ is known

Thus,

$$\begin{aligned}
0.95 &= \mathrm{P}\left(-1.96 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right) \\
&= \mathrm{P}\left(-1.96\frac{\sigma}{\sqrt{n}} \le \overline{X}-\mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) \\
&= \mathrm{P}\left(1.96\frac{\sigma}{\sqrt{n}} \ge \mu-\overline{X} \ge -1.96\frac{\sigma}{\sqrt{n}}\right) \\
&= \mathrm{P}\left(\overline{X}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}+1.96\frac{\sigma}{\sqrt{n}}\right)
\end{aligned}$$

## 95% CI for $\mu$ if $\sigma$ 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 $\mu$.

Using statistical jargon we say the random interval

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

covers $\mu$ 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 95% of observed samples the interval between 165cm and 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 $\mu$ if $\sigma$ is known

**Definition 8.** The $100(1-\alpha)$% CI for $\mu$ 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 $\mu$.

**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 $\mu$?

- Here $\overline{x} = 217$, $\sigma^2 = \sigma_0^2 = 46$ and $n = 12$.
- The 95% confidence interval the population mean $\mu$ is then

$$\begin{aligned}\overline{x} \pm 1.96 \times \frac{\sigma_0}{\sqrt{n}} &= 217 \pm 1.96 \times \frac{\sqrt{46}}{\sqrt{12}} \\ &= 217 \pm 3.84 \\ &= (213.16, 220.84)\end{aligned}$$

Tuesday, 16 October 2012

# Lecture 8 - Content

- **Confidence intervals continued**

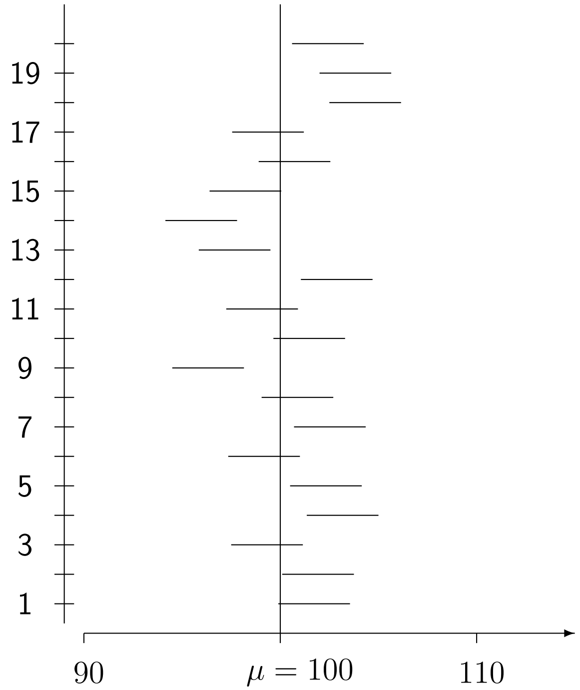
## Confidence intervals (cont)

### Properties of CIs

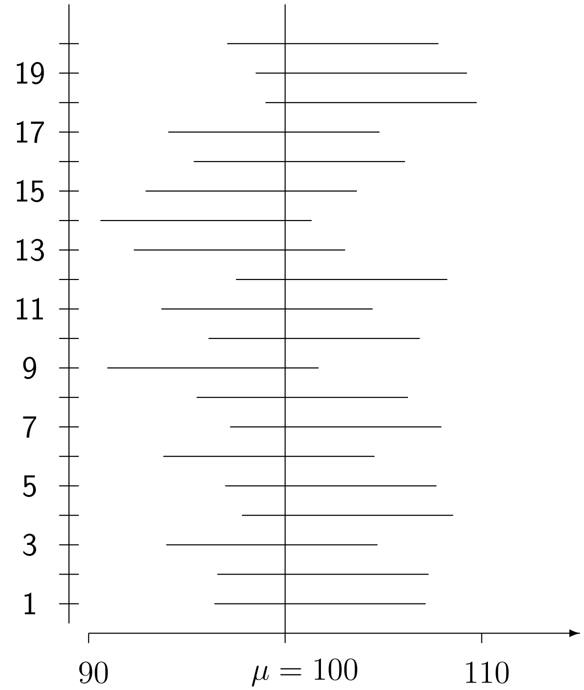
- Cover the true $\mu$ value with relative frequency approximately $(1-\alpha)$;
- as you increase $n$ the CI gets narrower;
- as you increase the confidence level, i.e. make $(1-\alpha)$ larger, the CI gets wider.

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

**50%-CI:** $\overline{X}_{30} \pm 0.67 \cdot \frac{15}{\sqrt{30}}$

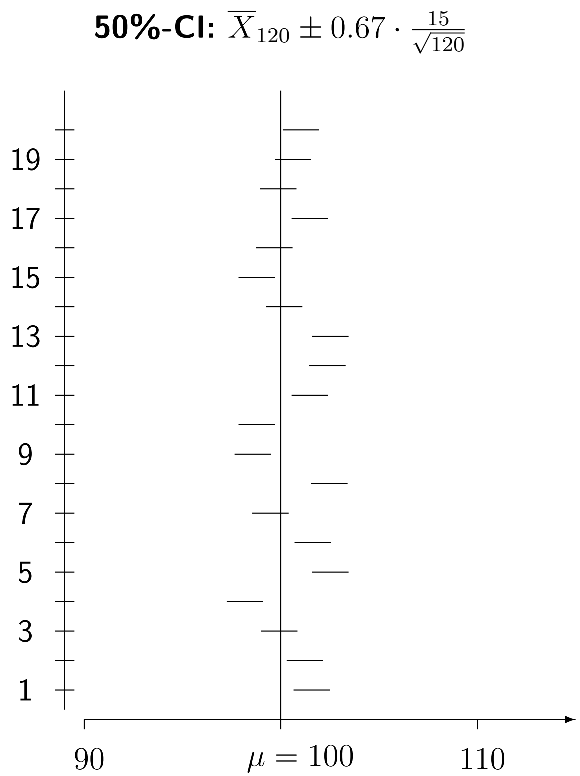


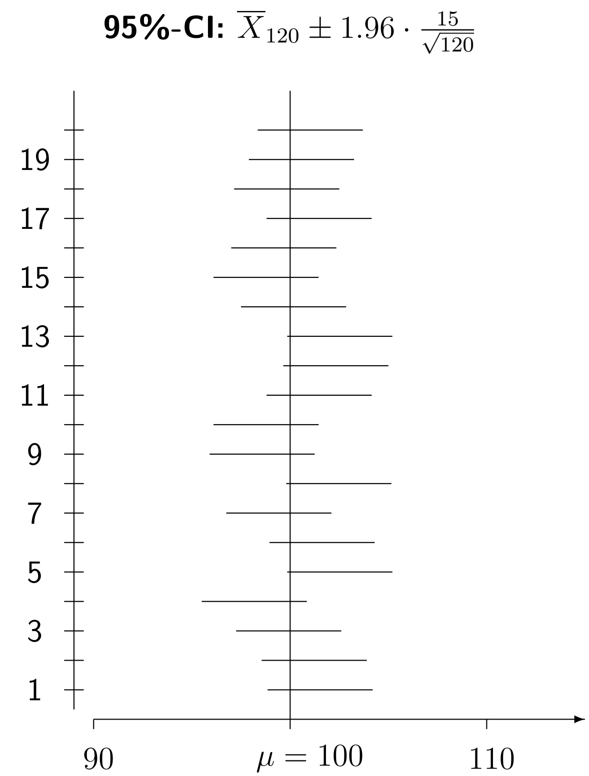

**95%-CI:** $\overline{X}_{30} \pm 1.96 \cdot \frac{15}{\sqrt{30}}$




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 $\mu$: 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 $\mu$: $0.99 = \mathrm{P}(-z_1 \leq Z \leq z_1) \Rightarrow z_1 = 2.576$ 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 $\mu$ if $\sigma$ 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 9.** A $100(1-\alpha)\%$ CI for $\mu$ of a normal population with unknown variance $\sigma^2$ 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 $\mu$?

- Here $\overline{x} = 217$, $s^2 = 46$ and $n = 12$.
- The 95% confidence interval the population mean $\mu$ is then

$$\overline{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

where solving $P(t_{11} > t^*) = 0.025$ via

$$t^* = \text{qt(0.975,11)} = 2.2 \quad \text{(to 2 d.p)}$$

- Then

$$\begin{aligned} \overline{x} \pm t^* \times \frac{s}{\sqrt{n}} &= 217 \pm 2.2 \times \frac{\sqrt{46}}{\sqrt{12}} \\ &= 217 \pm 4.31 \\ &= (212.69, 221.31) \end{aligned}$$

- 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 $\sigma$ 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 $\mu$ 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 10.** 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 $\pm 0.03$? [Note that as a convention, width is the same as the length of the CI, i.e. $\pm 0.03$ corresponds to a length = width = 0.06]

- Using the conservative estimate we want

$$1.96\sqrt{\frac{1}{4n}} \leq 0.03 \;\Rightarrow\; \frac{1.96}{2 \times 0.03} \leq \sqrt{n} \;\Rightarrow\; n \geq (32.6)^2 = 1067.1.$$

- Thus a sample of size 1068 is needed.

**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 $p$.

- ☐ Let $X$ be the number of functioning bulbs with $n = 1000$.
- ☐ Then $X \sim \mathcal{B}(1000, p)$ with $p$ unknown.
- ☐ If $T = 810$ we can estimate $p$ by $\widehat{p} = 810/1000$ and

$$p \sim N\left(\widehat{p}, \frac{\widehat{p}(1-\widehat{p})}{n}\right)$$

and a 95% confidence interval for $p$ is

$$\begin{aligned}\widehat{p} \pm 1.96 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} &= 0.81 \pm 1.96 \times \sqrt{\frac{0.81 \times (1-0.81)}{1000}} \\ &= 0.81 \pm 0.02 \\ &= (0.79, 0.83).\end{aligned}$$

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

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

Solve

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

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 $\alpha$ is typically 5%.