Confidence Intervals and Statistical testing

(Week 03 lecture notes)

Wei Q. Deng

Department of Statistical Sciences

July 16th 2018

Recap from last two lectures

Given a random sample $X_1, \ldots, X_n$ from some unknown distribution with some parameter of interest θ:

- Point estimator $\hat{\theta}$
  - method of moments
  - maximum likelihood estimator
  - Bias $\mathrm{E}(\hat{\theta} - \theta)$
  - Variance $\mathrm{Var}(\hat{\theta})$
  - Mean squared error
- Interval estimator (l, u)
  - $\mathrm{P}(l < \theta < u) = 1 - \alpha$
  - u and l should be functions of $\hat{\theta}$, but not functions of θ
  - sampling distribution of the estimator of $\hat{\theta}$ should involve θ

Interval estimators for the following parameters

- $\bar{X}$ and S^2 for μ and σ^2 for a normal random sample $X_i \sim \mathcal{N}(\mu, \sigma^2)$
- $\bar{X}$ for μ mean of a random sample (invoke the central limit theorem)
- p the population proportion

Topics today

- Interval estimator
    - Confidence interval for population porportion p (Binomial/Bernoulli)
    - Bootstrap confidence intervals
- Concepts of a statistical test
    - Structure of a test
    - Test statistics
    - Rejection region and P-values
    - One-sided vs. two-sided test
- Examples of statistical tests
    - Statistical tests of μ (normal, t, and binomial)
    - Statistical tests to compare two samples
        - test for equality of means
        - test for equality of variance (F-test)
        - assumptions for t-tests
        - A/B testing (e.g. for Average Revenue Per Paying User)

CI for p the proportion of success

- Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with $B(1, p)$. Denote $Y = \sum_{i=1}^{n} X_i \sim B(n, p)$,
    - A discrete probability distribution
    - Suppose each iid $X_i \in \{0, 1\}$ with probability of $X_i = 1$ being p, then $Y = \sum_{i=1}^{n} X_i$
    - Y: the number of successes in a sequence of n independent experiments
    - $Y \sim B(n, p)$
- First two moments of Y
    - $\mathrm{E}(Y) = np$ the number of expected successes
    - $\mathrm{Var}(Y) = np(1 - p)$
- First two moments of X_i
    - $\mathrm{E}(X_i) = p$
    - $\mathrm{Var}(X_i) = p(1 - p)$

Could you try the following using material introduced in lecture 2 and 3?

- Find an estimator of p using method of moments and maximum likelihood.
- Use CLT to conclude the sampling distribution of p
- Find the sampling distribution involving $\hat{p}$ and p
- Find the 95% CI for p

Method of moment

Let's match the first moment since we only have one unknown p.

$$m_1 = \sum_{i=1}^{n} X_i/n = \frac{Y}{n} = \mathsf{E}(X_i) = p$$

Maxmimum likelihood estimator

- Often the likelihood functions involve power or exponential function and it is much easier working with the log-likelihood instead.
- Know that $\text{argmax}_\theta L(\theta) = \text{argmax}_\theta \log L(\theta)$.

$$\frac{\partial \log L(\theta)}{\partial \theta} = \frac{1}{L(\theta)} \frac{\partial L(\theta)}{\partial \theta} = 0$$

- Start by writing out the likelihood function of the Bernoulli random variable sample with p.m.f. $(\theta = p)$.
- Then take the derivative with respect to p to get the result.

Step-by-step

Marginal probability of each X_i

$$\Pr(X = x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases}$$

The joint probability of $(X_1, \ldots, X_n)$:

$$\Pr(X_1, \ldots, X_n) = \Pi_{i=1}^n \Pr(X_i) = p^{\sum_{i=1}^n X_i}(1-p)^{n - \sum_{i=1}^n X_i}$$

The likelihood given data $(x_1, \ldots, x_n)$

$$L(p|(x_1, \ldots, x_n)) = p^{\sum_{i=1}^n x_i}(1-p)^{n - \sum_{i=1}^n x_i}$$

Take the log:

$$\log L(p) = (\sum_{i=1}^n x_i) \log p + (n - \sum_{i=1}^n x_i) \log (1-p)$$

Take the derivative with respect to p:

$$\frac{d \log L(p)}{dp} = (\sum_{i=1}^n x_i)\frac{1}{p} - (n - \sum_{i=1}^n x_i)\frac{1}{1-p} = 0$$

Solve for p:

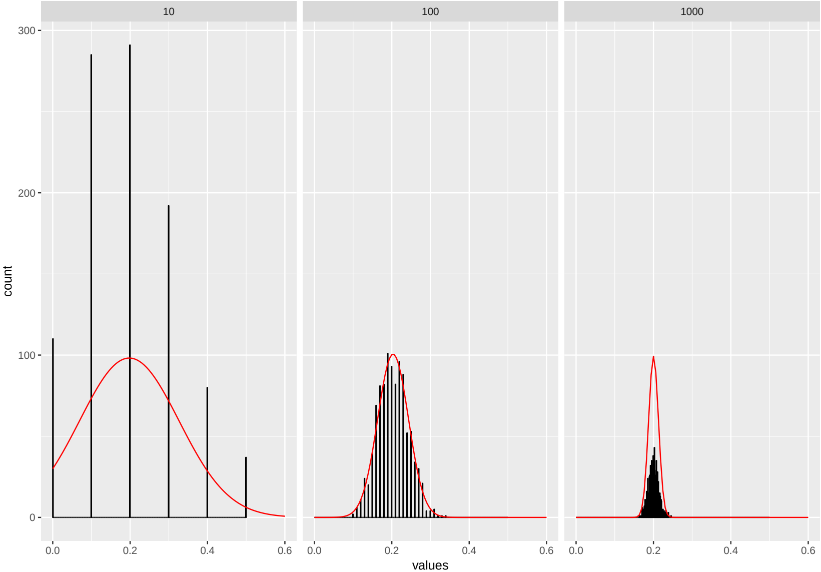
$$\hat{p} = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x} = \frac{y}{n}$$

An estimator of p

- The estimator from MoM and MLE is the same: $\hat{p} = \bar{X} = \frac{Y}{n}$.
- Verify that this estimator is unbiased (by showing $\mathrm{E}(\hat{p}) = p$).
- Can you come up with a sampling distribution for $\hat{p} = \frac{Y}{n}$?

Large sample behaviour

$\bar{X}$ is the mean of the sample and by CLT, if $np > 20$, approximately follows a normal distribution with mean p and variance $p(1-p)/n$.



With this information, we can proceed

- Estimator $\hat{p} = \frac{Y}{n}$
- Mean of estimator $\mathrm{E}(\hat{p}) = p$ (unbiased)
- Standard deviation of estimator $\sqrt{\mathrm{Var}(\hat{p})} = \sqrt{p(1-p)/n}$
- Define statistic $Q = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \to \mathcal{N}(0,1)$

This part is not tested:

- However, p in the denominator is uknown. We can solve a quadratic equation when locating u and l

$$\mathrm{P}(l < p < u) = \mathrm{P}\Big(\frac{\hat{p} - l}{\sqrt{p(1-p)/n}} > Q > \frac{\hat{p} - u}{\sqrt{p(1-p)/n}}\Big) = 1 - \alpha$$

  - We know that $\frac{\hat{p}-l}{\sqrt{p(1-p)/n}} = \Phi^{-1}(\alpha/2)$
  - $-\Phi^{-1}(\alpha/2) > Q > \Phi^{-1}(\alpha/2)$
  - Solving the inequality for p we have

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

An approximation to the large sample behaviour

- But, we see that if n is really large, then $\Phi^{-1}(\alpha/2)^2/2n \sim 0$ and $\Phi^{-1}(\alpha/2)^2/n \sim 0$, as well as $\Phi^{-1}(\alpha/2)^2/4n^2 \sim 0$.
- The interval then is approximately

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

- Essentially, we pretend that the variance is $\hat{p}(1-\hat{p})/n$ instead of $p(1-p)/n$.

Key insights about CIs

- When constructing a CI, we need to find the sampling distribution or the approximated sampling distribution of the estimator $\hat{\theta}$ relative to the parameter θ.
- When the sampling distribution is not easy to work with, we should try to find a quantity that contains both $\hat{\theta}$ and θ that does have a simple distribution that can be translated to quantiles (e.g. Z or T).
- For $\bar{X}$ from non-normal samples, we can invoke the CLT when $n > 30$ or $np > 20$, the CI will be approximated rather than exact.
- Know relationships among α (confidence level), width of the CI, sample size (n).

Open question

Combine what you now know about the bootstrap method and confidence interval, can you come up with an approximated 95% CI for the median of a random sample (any random sample) of size $n = 100$?

Bootstrap Confidence interval

Use the bootstrap samples obtained as previously described, the 95% bootstrap CI can be obtained by locating the $\alpha/2$ and $1-\alpha/2$ quantiles of the B bootstrap sample statistics $T(\hat{\theta}_1^*), \ldots, T(\hat{\theta}_B^*)$.

Concept of a statistical test

- Structure of a test
- Test statistics
- Rejection region and P-values
- One-sided vs. two-sided test

What is a statistical test?

- does the estimate from the sample agree with a known population value?
- provides a mechanism for making quantitative decisions
- determines whether there is sufficient evidence to “reject” a hypothesis (or what we believed to be the truth) about the process
  - Not rejecting implies we are willing to continue to **act as if we believe** the hypothesis is true.
  - Rejecting indicates we may not yet have enough data to **justify what we believed to be true**.

Consider the following example

- Let p be the proportion of defective circuit boards among all circuit boards produced by a certain manufacturer. Ideally, the proportion should be lower than 0.05.
- However, the total number of circuit boards manufactured could be well over 100,000. To quickly provide evidence for or against the claim that the defective rate is under 0.05, we need to rely on the power of **random sampling** and **hypothesis testing**.
- We can randomly sample 500 items, and these 500 can be checked individually for defects. Suppose the defective rate in this random sample is $\hat{p} = 0.02$.
- We can **test** whether $p = 0.05$ or $p \neq 0.05$ using information from this sample, and thus conclude whether the entire batch passed the quality control or not.

The essential components of a statistical test

- A random sample $(X_1, \ldots, X_n)$ of size **n** from a population of interest $f_{X|\theta}(x)$
  - For now, we restrict ourselves to data sampled from distributions that are of parametric families
  - The truth is then about whether certain values of θ is supported by the data or not
- A **pair** of hypotheses about the true or alternative values of θ
  - Null hypothesis, usually denoted by H_o
  - Alternative hypothesis, usually denoted by H_a or H_1
  - The null and alternative should be contradictory (i.e. no overlap)

The essential components of a statistical test (cont'd)

- $T(X)$: a test statistic, which we use to *test* against or for the hypothesis.
- P-value: a conditional probability that the observed test statistic is more extreme than would be expected under the null hypothesis.
- A significance level (α): an established level of risk for making a mistake (if the null hypothesis was wrong but we accepted it based on the evidence from data)
    - If P-value $< \alpha$, we **reject** the null hypothesis in favour of the alternative hypothesis.
    - Otherwise, we **fail** to reject the null hypothesis as the data do not provide sufficient evidence against it.

Hypothesis

- The **Null hypothesis**, denoted by H_o, is something we set out to believe is true.
  - It could be a convention or a commonly believed truth.
  - For example, when flipping a fair coin, we believe the probability of a head $p = 0.5$.
- The **Alternative hypothesis**, denoted by H_1 or H_a, is something contradictory to H_o.
  - It could be where we think the alternative truth lies.
  - For example, if we are interested in the coin being unfair, we can set the alternative to be $H_1 : p \neq 0.5$.
  - Or, if we are interested in the probability of a head being higher than the tail, the alternative hypothesis could be $H_1 : p > 0.5$.
- Notice that the hypotheses are always statements about the parameters. But how do we go about assessing which parameter values are more likely given the data?

Test statistics

Test statistics are also random variables and thus subject to randomness due to the sampling process.

- We need the sampling distribution of the test statistics under the null hypothesis (i.e. substituting the parameter value from the null hypothesis).
- Recall the interval estimators u and l, such that $\mathrm{P}(l < \theta < u) = 1 - \alpha$.
    - Suppose we have a statistic $T(X)$, and its sampling distribution involves θ and $\hat{\theta}$?
    - To obtain the interval estimator u and l, we had to invert the unknown θ to resemble T.
    - Now we are interested in rejection region instead of confidence intervals.
    - But suppose we are now given θ (according to the null hypothesis), and instead, we want to know what $\mathrm{P}(l > \theta)$ or $\mathrm{P}(\theta > u)$ is?
- This sampling distribution serves as the basis for all hypothesis testing.

An example with defective circuit boards

The defective rate of manufactured circuit board might be based on examining a random sample of $n = 200$ boards. Ideally, we hope the probability of a single board being defective to be lower than 0.1, the current quality control standard.

- Let X be the number of defective circuit boards in $n = 200$.
- Null hypothesis: $p = 0.1$
- Alternative hypothesis $p < 0.1$
- We expect there to be $\mathrm{E}(X) = np = 20$ defective circuit boards, if $X < 20$, then we could possibly reject the null hypothesis.
- If we reject then $X < 10$, then we have stronger evidence that $p < 0.1$.

Rejection region

Definition: The sample space of X where we have grounds to reject the null hypothesis.

- If we reject at $X < 20$, the rejection region is $\{0, 1, \ldots, 20\}$
- If we reject at $X < 10$, the rejection region is $\{0, 1, \ldots, 10\}$
- The size of the rejection region depends on how strict or liberal the testing procedure is (significance).
- Essentially, by setting the rejection region, we are setting the u or l values for which $\mathrm{P}(l > \theta)$ or $\mathrm{P}(\theta > u)$ is to be calculated.

Let's try an example with normal

Let $X_1, \ldots, X_n$ denote a random sample from a normal population distribution with σ^2 known.

- If we test the null hypothesis $\mu = \mu_o$ against $H_1 : \mu < \mu_o$ using the test statistic $\bar{X}$, show the rejection region of $\bar{X} < \mu_o - 2.33\sigma/\sqrt{n}$ has significance level $\alpha = 0.01$.
    - Identify the null hypothesis is about the mean parameter
    - Know that $\bar{X}$ is an estimator of μ
    - Know that $\bar{X}$ has a sampling distribution that depends on μ, i.e.

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$

    - The rejection region is defined by $\{X : Z < \Phi^{-1}(\alpha)\}$
    - But the rejection region needs to be at a certain level of significance

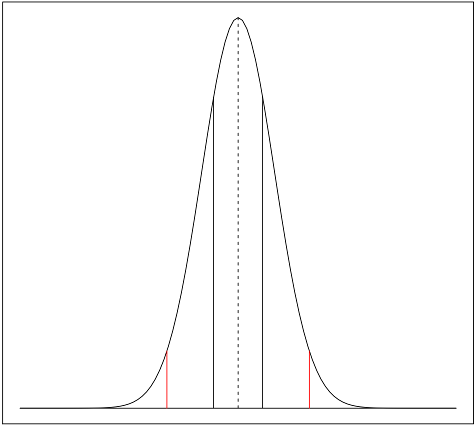
$$\mathrm{P}_X(Z < \Phi^{-1}(\alpha)) = \mathrm{P}(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \Phi^{-1}(\alpha)) = \alpha$$

    - Again, since $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$
    - $\frac{\bar{X}-\mu_o}{\sigma/\sqrt{n}} < -2.33$ under the null hypothesis

One-sided vs. two-sided test

- We have seen null hypothesis, and typically it is set at a fixed value: $\mu = \mu_o$ or $p = 0.5$, etc.
- There are two main types of alternative hypotheses:
  - Two-sided: $\mu \neq \mu_o$
  - One-sided: $\mu > \mu_o$ or $\mu < \mu_o$
- The rejection regions are clearly different depending on the alternative hypothesis:
  - Two-sided: $\{X : \mathrm{P}_X(T > t_{extreme} \text{ or } T < -t_{extreme}) = \alpha\}$
  - One-sided: $\{X : \mathrm{P}_X(T > t_{extreme}) = \alpha\}$ or $\{X : \mathrm{P}_X(T < -t_{extreme}) = \alpha\}$

Sampling distribution of $\bar{X}$



Statistical tests of μ (normal, t, and binomial)

One sample tests about a population mean

Let $X_1, \ldots, X_n$ be a **normal random sample** of size n with *mean* μ and variance σ^2. The mean estimator is $\bar{X}$ and consider the test statistics:

- $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$
- $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$, where $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$

Let $X_1, \ldots, X_n$ be a **random sample** of size $n > 30$ with *mean* μ and variance σ^2. The mean estimator is $\bar{X}$ and consider the test statistics:

- $Z' = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ (approximated)
- $T' = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$, where $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$

Let $X_1, \ldots, X_n \in \{0,1\}$ be a **random sample** of size n. The *proportion* estimator is $\hat{p} = \sum_{i=1}^{n} X_i/n$, $n\hat{p} > 20$ and consider the test statistic:

- $Y = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim \mathcal{N}(0,1)$ (approximated)

Testing about μ or p

In the cases above, we can test the null hypotheses $\mu = \mu_o$ or $p = p_o$ by using the sampling distribution of Z, T, and Y, for which the quantiles can be easily calculated.

Define the p-value of a test given the test statistic Z (random variable) of a two-sided test:

$$\text{p-value} = 2\text{min}(\text{P}(Z < -z_{\text{obs}} | \mu = \mu_o), \text{P}(Z > z_{\text{obs}} | \mu = \mu_o))$$

Define the p-value of a test given the test statistic Z (random variable) of a one-sided test:

$$\text{p-value} = \text{P}(Z > z_{\text{obs}} | H_o) \text{ when null is } \mu < \mu_o$$

or

$$\text{p-value} = \text{P}(Z < z_{\text{obs}} | H_o) \text{ when null is } \mu > \mu_o$$

where, for example, $z_{\text{obs}} = \frac{\bar{x} - \mu_o}{\sigma / \sqrt{n}}$.

An example: House Insulation: Whiteside's Data

Mr Derek Whiteside of the UK Building Research Station recorded the weekly gas consumption and average external temperature at his own house in south-east England for two heating seasons, one of 26 weeks before, and one of 30 weeks after cavity-wall insulation was installed. Is the **average external temperature** at his house the same as the **national average** of **5.9** celcius in Januaray?

```
##    Insul Temp Gas
## 1 Before -0.8 7.2
## 2 Before -0.7 6.9
## 3 Before  0.4 6.4
## 4 Before  2.5 6.0
## 5 Before  2.9 5.8
## 6 Before  3.2 5.8
```

An example: House Insulation: Whiteside's Data

```
library(MASS)
data(whiteside)
c(length(whiteside$Temp), mean(whiteside$Temp), sd(whiteside$Temp))
```

```
## [1] 56.000000  4.875000  2.749562
```

```
import pandas
import numpy as np
whiteside = pandas.read_csv("whiteside.csv")
n = len(whiteside[["Temp"]])
x = np.array(whiteside[["Temp"]])
print(n)
```

```
## 56
```

```
print(x.mean())
```

```
## 4.875
```

```
print(np.std(x, ddof=1))
```

```
## 2.74956194858
```

Breaking it down:

- What is the random sample?
  - The external temperature at his house $(X_1, \ldots, X_{56})$
- What is the null hypothesis?
  - The national average temperature μ_o is the same as the average at his house (μ truth about his house): $\mu = \mu_o = 5.9$
- What is the test statistic?
  - Assuming large sample, invoke the CLT to conclude that $\bar{X}$ is normally distributed
  - But since the variance is unknown, we substitute for the sample variance.
  - $T' = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$, where $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$
- What is the observed value of the test statistic under the null hypothesis?
  - $t' = \frac{\bar{x}-\mu}{s/\sqrt{n}} = -2.789$
  - $\bar{x} = 4.875$
  - $s = \sqrt{s^2} = 2.750$
- What is the p-value for evidence against the null?
  - $\mathrm{p} = 2\mathrm{P}(T' < -2.789) = 0.0072$
  - Since $\mathrm{p} < \alpha = 0.01$, we say we reject the null hypothesis of equality at 1% significance level in favour of the alternative hypothesis.

Statistical tests to compare two samples

- test for equality of means
  - independent two samples t-tests
  - paired t-tests
- assumptions for t-tests
- test for equality of variance (F-test)

test for equality of means for independent samples

- So far we looked at μ or p with respect to a single sample, and tested it compared to a population.
- What about two samples?
- Let $X_1, \ldots, X_n$ and $X_1', \ldots, X_{n'}'$ be two random samples, for example, the height for a group of n female high school students and a group of n' male high school students.
- We are interested in whether the mean height in female and male differ, i.e. $\mu_F = \mu_M$ is a natural choice for our null hypothesis.
- But we will see a few slides later that the null should be subtly changed to $\mu_F - \mu_M = 0$.

Basic setup of two sample problem for independent samples

- Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample with mean μ_1 and variance σ_1^2
- Let $\mathbf{X}' = (X_1', \ldots, X_{n'}')$ be a random sample with mean μ_2 and variance σ_2^2
- The two samples $\mathbf{X}$ and $\mathbf{X}'$ are independent.

Notice that

- n does not necessarily have to be the same as n', possibly allowing different sample sizes.
- σ_1^2 does not necessarily have to be the same as σ_2^2, possibly allowing different true variance values.
- the goal is to test $H_o : \mu_1 = \mu_2$ against
- $H_1 : \mu_1 \neq \mu_2$, or
- $H_1 : \mu_1 > \mu_2$ or
- $H_1 : \mu_1 < \mu_2$

Going back to the hypothesis test

- What is the parameter we are testing?
  - $d = \mu_1 - \mu_2 = 0$
- What is an estimator for the parameter d (show that it is unbiased)?
  - $\hat{d} = \bar{X} - \bar{X'}$
- What is the variance of this estimator?
  - $\sigma_d^2 = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{X'}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n'}$
- What is the sampling distribution of $\bar{X} - \bar{X'}$?
  - We can write out the $Z_d = \frac{(\bar{X} - \bar{X'}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n'}}}$

Sampling distribution of Z_d

$$Z_d = \frac{(\bar{X} - \bar{X}') - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n'}}}$$

- If the two samples are both normal, then $Z_d \sim \mathcal{N}(0, 1)$.
- If the sample sizes $n > 30$ and $n' > 30$, by CLT, both $\bar{X}$ and $\bar{X}'$ are then approximately normal, so $Z_d \to \mathcal{N}(0, 1)$.

What about when the variances are unknown?

We have to estimate σ_1^2 and σ_2^2:

- Equal sample size, equal variance: If $n_1 = n_2 = n$ and we can believe that $\sigma_1^2 = \sigma_2^2$, then define $S_p = \sqrt{(S_1^2 + S_2^2)/2}$:

$$T = \frac{(\bar{X} - \bar{X}\prime) - (\mu_1 - \mu_2)}{S_p\sqrt{2/n}} \sim t(2n-2)$$

- Unequal sample size, equal variance: If $n_1 \neq n_2$ and we can believe that $\sigma_1^2 = \sigma_2^2$, then define $S_p = \sqrt{((n_1 - 1)S_1^2 + (n_2 - 1)S_2^2)/(n_1 + n_2 - 2)}$:

$$T = \frac{(\bar{X} - \bar{X}\prime) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2)$$

- Unequal variance: if $\sigma_1^2 \neq \sigma_2^2$, then they have to estimated separately,

$$T = \frac{(\bar{X} - \bar{X}\prime) - (\mu_1 - \mu_2)}{\sqrt{\frac{\S_1^2}{n} + \frac{S_2^2}{n'}}} \sim t(df)$$

approximately, where
$df = \left(s_1^2/n_1 + s_2^2/n_2\right)^2 / \left((s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)\right).$

Test for equality of variance (F-test)

For a formal testing of the two variances σ_1^2 and σ_2^2, we use an F-test. It will be introduced with more details after the midterm.
For now, just know how to perform this test in R/Python and interpret the results.

- Two independent **normal** samples each with variance σ_1^2 and σ_2^2, and sample size n and n'.
- Null Hypothesis: $\sigma_1^2 = \sigma_2^2$
- Alternative Hypothesis: $\sigma_1^2 \neq \sigma_2^2$
- Test statistics: $F = \frac{S_1^2}{S_2^2} \sim F(n-1, n'-1)$
- If p-value less than α, then we reject the null of variance equality and use Welch's t-test.
- Otherwise, conclude that the pooled variance estimate is sufficient and use the original two sample t-test assuming equal variance.

An example: House Insulation: Whiteside's Data

Mr Derek Whiteside of the UK Building Research Station recorded the weekly gas consumption and average external temperature at his own house in south-east England for two heating seasons, one of 26 weeks before, and one of 30 weeks after cavity-wall insulation was installed. Has the **average external temperature** at his house changed after the insulation was install?

```r
library(MASS)
data(whiteside)
tapply(whiteside$Temp, whiteside$Insul, length)
```

```
## Before  After
##     26     30
```

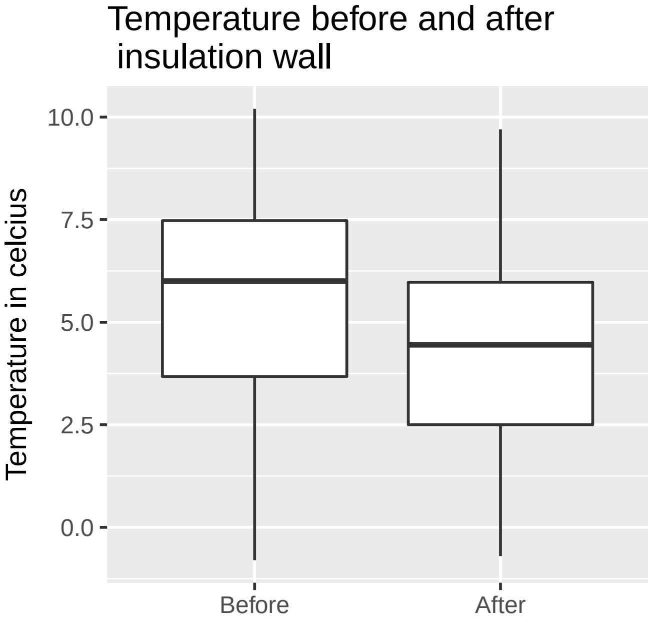
```r
tapply(whiteside$Temp, whiteside$Insul, mean)
```

```
##   Before    After
## 5.350000 4.463333
```

```r
tapply(whiteside$Temp, whiteside$Insul, sd)
```

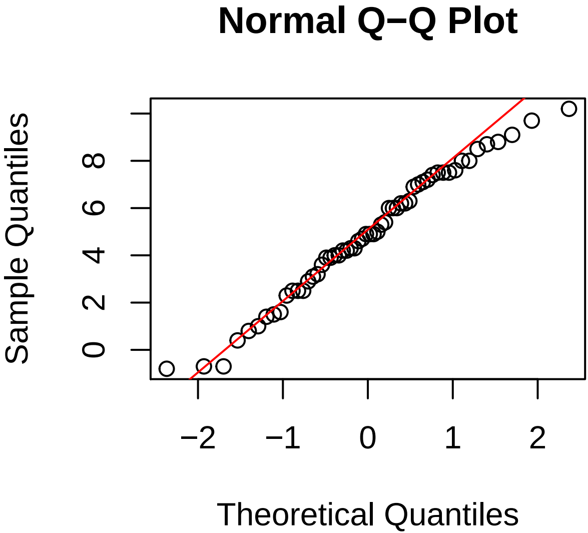
```
##   Before    After
## 2.872804 2.616458
```

Graphical illustration of the data



Checking normality - Quantile-quantile plots

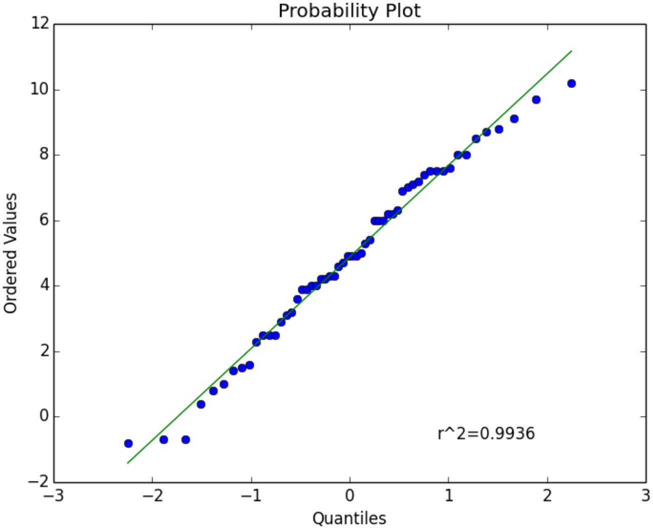
```
qqnorm(whiteside$Temp)
qqline(whiteside$Temp, col=2)
```



In Python:

```
import pandas
import numpy as np
import scipy.stats as stats
import pylab
import matplotlib.pyplot as plt
whiteside = pandas.read_csv("whiteside.csv")
x=whiteside[["Temp"]]
x = np.concatenate(x)
stats.probplot(x, dist="norm", plot=pylab)
pylab.show()
```

In Python:


Probability Plot
Ordered Values
Quantiles
12
10
8
6
4
2
0
−2
−3
−2
−1
0
1
2
3
r^2=0.9936

In R:

```r
# test variance equality
var.test(whiteside$Temp~whiteside$Insul)
```

```
##
##  F test to compare two variances
##
## data:  whiteside$Temp by whiteside$Insul
## F = 1.2055, num df = 25, denom df = 29, p-value = 0.624
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5628432 2.6403094
## sample estimates:
## ratio of variances
##           1.205548
```

```r
library(car)
leveneTest(whiteside$Temp~whiteside$Insul)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.1273 0.7226
##       54
```

In R (cont'd)

```r
t.test(whiteside$Temp~whiteside$Insul)
```

```
##
##  Welch Two Sample t-test
##
## data:  whiteside$Temp by whiteside$Insul
## t = 1.2004, df = 51.099, p-value = 0.2355
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5961847  2.3695180
## sample estimates:
## mean in group Before  mean in group After
##             5.350000             4.463333
```

```r
t.test(whiteside$Temp~whiteside$Insul, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  whiteside$Temp by whiteside$Insul
## t = 1.2085, df = 54, p-value = 0.2321
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5842485  2.3575819
## sample estimates:
## mean in group Before  mean in group After
##             5.350000             4.463333
```

In Python:

```
import pandas
import numpy as np
from scipy import stats
whiteside = pandas.read_csv("whiteside.csv")
before = whiteside[whiteside['Insul'] == 'Before']['Temp']
after = whiteside[whiteside['Insul'] == 'After']['Temp']
print(stats.levene(before, after))
```

```
## (0.12730790476410858, 0.72263030808380435)
```

```
print(stats.ttest_ind(before, after))
```

```
## (array(1.2085398488128547), 0.23210531658694758)
```

```
print(stats.ttest_ind(before, after, equal_var=False))
```

```
## (array(1.2003723939884317), 0.23553013466958814)
```

test for equality of means for paired samples

What if the two samples are not independent?

- We have solution when the two samples are completely dependent, i.e. the observations are paired.
- What are paired observations?
  - Suppose for each patient, we collect two blood pressure readings, one in the morning and one in the afternoon.
  - Suppose for each merchandise, we record the number of sales before clearance and after clearance.
- Notice that the sample sizes must be exactly the same, and we are interested in the relative differences at each sample, rather than the overall difference between the two samples.

Examples

- We randomly select 20 males and 20 females and compare the average time they spend watching TV. Is this an independent sample or paired sample?
- We randomly select 20 couples and compare the time the husbands and wives spend watching TV. Is this an independent sample or paired sample?

Basic setup of two sample problem for paired samples

- Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample with mean μ_1 and variance σ_1^2
- Let $\mathbf{X}' = (X_1', \ldots, X_n')$ be a random sample with mean μ_2 and variance σ_2^2
- Define $\mathbf{D} = (D_1, \ldots, D_n)$, where $D_i = X_i - X_i'$.

What is the sampling distribution of $\bar{D}$?

- **D** is a random sample with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.
- This can be reduced to the one sample test depending on
  - whether σ_1^2 and σ_2^2 are known (normal or t test statistics)
  - whether the two samples are normally distributed (exact or large sample distribution)
- Let's assume the variances are not known. The test statistic has a t-distribution

$$T = \frac{\bar{D}}{S_D/\sqrt{n}} \sim t(n-1)$$

where $S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}$.

assumptions for t-tests

Independent two-sample t-test

- two samples are independent
- two variances are equal (use pooled variance estimate)
- samples are from normal (but relaxed for large samples)

Paired two-sample t-test

- two samples are paired
- each pair sample is independent from other paired samples
- difference between samples is normal (but relaxed for large samples)

Learning goals for two sample and paired t-tests

- Given a question, setup the null and alternative hypotheses correctly.
- Know the sampling distribution (Z or T) for a given problem
- Know the different versions of t-tests under different assumptions
- You should be able to perform these in R/Python and read outputs from these tests

A/B testing example

You are tasked to perform A/B testing on two versions of the on-line shopping website A and B. The web server has been setup to randomly split users to A or B as they click on the website. The two variations A and B are different only in webpage layouts. Your employer is interested in whether the webpage layouts influence the average revenue per paying user.

- Setup the null and alternative hypotheses.
- What is a reasonable distribution for the mean revenue?
- What is the test statistic?
- What assumptions do you think are reasonable to make?
- Can you derive the sampling distribution of the test statistic based on reasonable assumptions?
- What other considerations would you include in your proposal?

Good resources for downloading data

- Kaggle data competition
- UCI machine learning data respository
- MASS R library
- Google trends US