Lecture Three

Normal theory null distributions

Normal (Gaussian) distribution

The normal distribution is often relevant because of the *Central Limit Theorem* (CLT):

> *A random variable which is a sum of 'many' independent random variables will have an (approximately) normal distribution.*

Examples

(1) Many natural responses may be modelled as the additive effect of many factors.
e.g. crop yield:

$$\begin{aligned} y_1 &= a_1 x_{\text{seed}1} + a_2 x_{\text{soil}1} + a_3 x_{\text{water}1} + \cdots \\ y_2 &= a_1 x_{\text{seed}2} + a_2 x_{\text{soil}2} + a_3 x_{\text{water}2} + \cdots \\ \vdots \;& \vdots \;\; \vdots \\ y_n &= a_1 x_{\text{seed}n} + a_2 x_{\text{soil}n} + a_3 x_{\text{water}n} + \cdots \end{aligned}$$

where

$x_{\text{seed}1}, \ldots, x_{\text{seed}n}$ are independent samples from a (not necessarily normal) distribution with mean μ_{seed} and variance σ^2_{seed};

$x_{\text{soil}1}, \ldots, x_{\text{soil}n}$ are independent samples from a distribution with mean μ_{soil} and variance σ^2_{soil};

$x_{\text{water}1}, \ldots, x_{\text{water}n}$ are independent samples from a distribution with mean μ_{water} and variance σ^2_{water};

it then follows that $(y_1, \ldots, y_n)$ will be independent samples from an approximately normal joint distribution with

$$\begin{aligned} \mu_Y &= a_1 \mu_{\text{seed}} + a_2 \mu_{\text{soil}} + a_3 \mu_{\text{water}} + \cdots \\ \sigma^2_Y &= a_1^2 \sigma^2_{\text{seed}} + a_2^2 \sigma^2_{\text{soil}} + a_3^2 \sigma^2_{\text{water}} + \cdots \end{aligned}$$

additive effects $\Rightarrow$ normally distributed data

(2) The sampling distribution for $\bar{Y}$ from independent samples from a population

Recap: sampling distribution:

$$
\begin{array}{l}
\text{Population A} \Rightarrow \text{sample } (y_1^{(1)}, \ldots, y_n^{(1)}) \Rightarrow \bar{y}^{(1)} \\
\text{Population A} \Rightarrow \text{sample } (y_1^{(2)}, \ldots, y_n^{(2)}) \Rightarrow \bar{y}^{(2)} \\
\qquad\qquad \vdots \qquad\qquad\qquad \vdots \\
\text{Population A} \Rightarrow \text{sample } (y_1^{(N)}, \ldots, y_n^{(N)}) \Rightarrow \bar{y}^{(N)}
\end{array}
$$

Distribution of $(\bar{y}^{(1)}, \ldots, \bar{y}^{(N)})$ is called the *sampling distribution* of the sample mean $\bar{Y}$.

For 'reasonable' distributions (finite mean μ and variance σ^2) and non-tiny sample sizes ($n > 30$), $\bar{Y}$ will have an approximately normal distribution, with mean μ, variance σ^2/n.

Why do we care about the sampling distribution of $\bar{Y}$?

Consider H_0: $E(Y_A) = E(Y_B) = \mu$ (treatment has no effect)

Then *regardless of the distribution of the data*, under H_0 we have

$$\bar{Y}_A \dot{\sim} N(\mu, \sigma^2/n_A) \qquad \bar{Y}_B \dot{\sim} N(\mu, \sigma^2/n_B)$$

hence

$$\bar{Y}_A - \bar{Y}_B \dot{\sim} N(0, (\sigma^2/n_A) + (\sigma^2/n_B))$$

represents a distribution of hypothetical results of a sampling experiment that it could have occurred under H_0.

(Review of) properties of the Normal Distribution (I)

(1) If $Y \sim N(\mu, \sigma^2)$, then $aY + b \sim N(a\mu + b, a^2\sigma^2)$; in particular, $(Y - \mu)/\sigma \sim N(0, 1)$
$N(0, 1)$ is called *the standard Normal distribution*.

(2) If $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$ and Y_1, Y_2 independent then $Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

(3) If $Y_1, \ldots, Y_n$ are an i.i.d. sample from $N(\mu, \sigma^2)$ then $\bar{Y}$ is *statistically independent* of

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Normal theory Null Distributions

Consider the simple hypothesis test:

- H_0: $E(Y) = \mu_0$
- H_1: $E(Y) \neq \mu_0$ (two-sided)

obtain data $y_1, \ldots, y_n$, evaluate H_0, H_1 with test statistic:

$$d(\mathbf{y}) = d(y_1, \ldots, y_n) = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$$

1. If data are from a normal population $N(\mu, \sigma^2)$ then

$$\begin{aligned} \bar{Y} - \mu_0 &\sim N(\mu - \mu_0, \sigma^2/n) \\ d(\mathbf{Y}) = \frac{(\bar{Y} - \mu_0)}{(\sigma/\sqrt{n})} &\sim N(\frac{\mu - \mu_0}{\sigma/\sqrt{n}}, 1) \end{aligned}$$

   and thus if H_0 is true then $d(\mathbf{Y}) \sim N(0, 1)$.

   Thus if the null hypothesis is true then we would expect $d(\mathbf{Y})$ to have a standard normal distribution, e.g. in 95% of samples taking a value between -1.96 and $+1.96$.

   Equivalently: the null distribution is standard normal.

2. If data are not normal, but H_0 is true, the variance is σ^2, and n is fairly big (e.g. $n > 30$), then by the CLT we have:

$$d(\mathbf{Y}) = \frac{(\bar{Y} - \mu_0)}{(\sigma/\sqrt{n})} \dot{\sim} N(0, 1)$$

Hypothesis Test

Large values of $|d(\mathbf{Y})|$ provide strong evidence against H_0 $\Rightarrow$ Reject H_0 for larger values of $|d(\mathbf{Y})|$

$$\begin{aligned} \text{p-value} = Pr(|d(\mathbf{Y})| \geq |d(\mathbf{y}_{\text{obs}})|) &= Pr(Z \leq -|d(\mathbf{y}_{\text{obs}})|) + Pr(Z \geq |d(\mathbf{y}_{\text{obs}})|) \\ &= 2Pr(Z \geq |d(\mathbf{y}_{\text{obs}})|) \\ &= 2 * (1 - \texttt{pnorm}(|d(\mathbf{y}_{\text{obs}})|, 0, 1)) \end{aligned}$$

(here we use Z to indicate a standard normal RV).
Problem: σ^2 is usually unknown

Thus we cannot compute $d(\mathbf{Y})$ (in this sense it is not a genuine test-statistic).

Solution: approximate σ^2 with s^2, the sample variance, and use

$$t(\mathbf{y}) = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

Q: What is the distribution of $t(\mathbf{Y})$ under H_0: $E(Y) = \mu_0$?
Well, $s \to \sigma$, hence we might hope that:

$$\frac{\bar{y} - \mu_0}{s/\sqrt{n}} \approx \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

For very large samples (> 100) this is a reasonable approximation, but for less large samples we need to take into account that S varies around σ.

Properties of the Normal distribution (2):

χ^2 and t-distributions

(4) If $Z_1, \ldots, Z_n \sim_{\text{i.i.d.}} N(0,1)$ then

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

the χ^2 ('Chi-squared') distribution with n d.f. (degrees of freedom)

In particular, if $Y_1, \ldots, Y_n \sim$ i.i.d.$N(\mu, \sigma^2)$, then

$$\begin{aligned} \frac{Y_1 - \mu}{\sigma}, \frac{Y_2 - \mu}{\sigma}, \ldots \frac{Y_n - \mu}{\sigma} \quad &\sim_{\text{i.i.d.}} \quad N(0,1) \\ \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2 \quad &\sim \quad \chi_n^2 \\ \frac{n-1}{\sigma^2}S^2 = \frac{n-1}{\sigma^2}\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \quad &\sim \quad \chi_{n-1}^2 \end{aligned}$$

Note: the d.f. is the number of *independent* normal RVs that are summed and squared, or equivalently, the dimension of the space in which these variables live:

- Clearly $\frac{Y_i - \mu}{\sigma}$ is independent of $\frac{Y_j - \mu}{\sigma}$;
- But $\frac{Y_j - \bar{Y}}{\sigma}$ is *not* independent of $\frac{Y_{j^*} - \bar{Y}}{\sigma}$, for $j \neq j^*$.
- However, if we define $W_i = Y_i - \bar{Y}$ then

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n} W_i^2 = \frac{1}{\sigma^2}\sum_{i=1}^{\mathbf{n-1}} U_i^2$$

where for $i = 1, \ldots, n-1$:

$$U_i = -\sqrt{\frac{i}{i+1}} W_{i+1} + \frac{1}{\sqrt{i(i+1)}} \sum_{j=1}^{i} W_j$$

and U_j, U_{j^*} are independent for $j \neq j^*$.
Thus we see that $n-1$ is indeed the correct d.f. for S^2.

(5) If $Z \sim N(0,1)$, $X \sim \chi^2_k$ and Z and X are independent then

$$\frac{Z}{\sqrt{X/k}} \sim t_k$$

Student's t-distribution on k d.f. ('Student' was a pseudonym used by W. Gosset)
Note that as $k \to \infty$ $t_k \to N(0,1)$ distribution.

The t-distribution has *heavier tails* than the standard normal, i.e. for large values of x:

$$P(|Z| > x) < P(|T| > x)$$

where $Z \sim N(0,1)$ and $T \sim t_k$.

Back to the hypothesis test:

We now return to the question: if $E(Y) = \mu_0$, so H_0 is true, and in addition, $Y_1, \ldots Y_n \sim_{\text{i.i.d.}} N(\mu_0, \sigma^2)$, what is the distribution of

$$t(\mathbf{Y}) = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}?$$

- We already showed that if $Y_1, \ldots Y_n \sim_{\text{i.i.d.}} N(\mu_0, \sigma^2)$

$$\frac{(\bar{Y} - \mu_0)}{(\sigma/\sqrt{n})} \sim N(0,1) \quad \text{``}Z''$$

- Further, from (4) it follows that

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1} \quad \text{``}X'' \text{ with ``}k'' = n-1$$

- Finally from (3) we have that $\bar{Y}$ and S^2 are independent. Hence "Z''" and "X''" are independent.

- Thus it follows that

$$\frac{\bar{Y} - \mu_0}{S/\sqrt{n}} = \frac{\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{n-1}{\sigma^2}S^2}{n-1}}} = \frac{\text{``}Z''}{\sqrt{\text{``}X''/(n-1)}} \sim t_{n-1}$$

- Note that neither "X''" nor "Z''" are test statistics, since (unless we know σ^2) they cannot be computed from the sample.

One sample t-test

- H_0: $E(Y) = \mu_0$
- H_1: $E(Y) \neq \mu_0$ (two-sided)
- Assumption: if H_0 is true then $Y_1, \ldots, Y_n \sim_{\text{i.i.d.}} N(\mu_0, \sigma^2)$
- Data $y_1, \ldots, y_n$
- Test statistic:

$$t_{\text{obs}} = t(\mathbf{y}) = \sqrt{n}\left(\frac{\bar{y} - \mu_0}{s}\right)$$

- Null distribution: from the assumptions, if H_0 is true then $t(\mathbf{Y}) \sim t_{n-1}$, i.e. the population of hypothetical values of $t(\mathbf{Y})$ that might have been sampled is a t-distribution on $n-1$ d.f.
- p-value: measuring evidence against H_0:

$$p = Pr\left(|t(\mathbf{Y})| \geq |t_{\text{obs}}| \,\middle|\, H_0\right) = 2 * (\mathtt{1} - \mathtt{pt}(|t_{\text{obs}}|, \mathtt{n} - \mathtt{1}))$$

  You can also try `t.test(data.vector, mu=`μ_0`)`

Basic Decision Theory

Goal: Reject H_0 when false, accept H_0 when true.

Method: Judge evidence provided by data against H_0.

	Truth	
Decision	H_0 true	H_0 false
Accept H_0:	Correct	Type II error
Reject H_0:	Type I error	Correct

Alternate terminology:

- H_0: ‘no treatment effect’ / ‘you don’t have the disease’
- H_1: ‘treatment effect’ / ’you do have the disease’

then Type I errors correspond to *False positives*;
Type II errors correspond to *False negatives*.
Note: A type I error can *only* occur when the null hypothesis is true.
Conversely, a type II error can *only* occur when the alternative is true.

p-value Decision Procedure:

(i) Select a level α $(0 < \alpha < 1)$

(ii) Compute the observed value of the statistic, and hence the *p-value*.

(iii) Reject H_0 if *p-value*$\leq \alpha$;
accept H_0 if *p-value*$> \alpha$.

Note: *A large p-value does* **not** *imply that the alternative hypothesis is false, merely that there is little evidence against the null hypothesis. For this reason some people speak of 'failing to reject'* H_0*, rather than 'accepting'* H_0*; the implication being that the null hypothesis has 'escaped' rejection this time, but might not be so 'lucky' next time.*

Example: One sample t-test

- H_0: $E(Y) = \mu_0$
- H_1: $E(Y) \neq \mu_0$ (two-sided)

Let t_{obs} be the observed value of the t-statistic $t(\mathbf{y}) = \sqrt{n}(\bar{y} - \mu_0)/s$. Let T_{n-1} be a random variable with a Student t-distribution on $n-1$ d.f.

$$\begin{array}{rrcl}
\text{reject } H_0 \text{ iff} & \text{p-value} & \leq & \alpha \\
\Leftrightarrow & Pr(|T_{n-1}| \geq |t_{\text{obs}}|) & \leq & \alpha \\
\Leftrightarrow & 2 \times Pr(T_{n-1} \geq |t_{\text{obs}}|) & \leq & \alpha \\
\Leftrightarrow & Pr(T_{n-1} \geq |t_{\text{obs}}|) & \leq & \alpha/2 \\
\Leftrightarrow & 1 - Pr(T_{n-1} \leq |t_{\text{obs}}|) & \leq & \alpha/2 \\
\Leftrightarrow & Pr(T_{n-1} \leq |t_{\text{obs}}|) & \geq & 1 - \alpha/2
\end{array}$$

Let $t_{n-1,1-\alpha/2}$ be the $1-\alpha/2$ quantile of the t-distribution with $n-1$ d.f. i.e. $Pr(T_{n-1} < t_{n-1,1-\alpha/2}) = 1 - \alpha/2$.

Let x_1, x_2, x_3 be three numbers such that

$$x_1 < -t_{n-1,1-\alpha/2} < x_2 < t_{n-1,1-\alpha/2} < x_3$$

(picture)

- If $t_{\text{obs}} = x_1$ then $Pr(T_{n-1} \leq |x_1|) > 1 - \alpha/2$, hence p-value $< \alpha$, so we *reject* H_0.
- If $t_{\text{obs}} = x_2$ then $Pr(T_{n-1} \leq |x_2|) < 1 - \alpha/2$, hence p-value $> \alpha$, so we *accept* H_0.
- If $t_{\text{obs}} = x_3$ then $Pr(T_{n-1} \leq |x_3|) < 1 - \alpha/2$, hence p-value $< \alpha$, so we *reject* H_0.

Thus we see that for a one sample t-test, the p-value decision procedure is equivalent to

t-statistic Decision Procedure:

(i) Select a level α $(0 < \alpha < 1)$

(ii) Compute the observed t-statistic $t_{\text{obs}} = t(\mathbf{y})$

(iii) Reject H_0 if $|t_{\text{obs}}| \geq t_{n-1,1-\alpha/2}$;
Accept H_0 if $|t_{\text{obs}}| < t_{n-1,1-\alpha/2}$

Terminology The range of values of the test statistic where (for a given α) the null hypothesis is rejected is called the *rejection region*; conversely the range of values for which the null is not rejected is called the *acceptance region*.
It should now be obvious that for both decision procedures that

$$\begin{aligned} P(\text{type I error} \mid H_0 \text{ is true }) &= P(\text{reject } H_0 \mid H_0 \text{ is true }) \\ &= P(|t(\mathbf{Y})| \geq t_{n-1,1-\alpha/2} \mid H_0 \text{ is true}) \\ &= 2 \times \alpha/2 = \alpha \end{aligned}$$

Such a procedure is called a *level-α test*.
α controls the *pre-experimental* type-I error rate.
Note that we can construct a decision procedure to give any specified type I error rate.

Note: (Interpretation)
If every experimenter used, for example, $\alpha = 0.05$ then...

- The null hypothesis would be falsely rejected for 5% of those experiments in which H_0 is true, and would correctly not be rejected for 95% of these experiments (where H_0 is true).

- What about those experiments where H_1 is true? For what proportion of these will we incorrectly accept H_0, and for what proportion will we correctly reject H_0? i.e. for such experiments what will our Type II error rate be?

- We will return to this question when we discuss *power* and the control of type II errors. (It will depend on *how* H_1 is true, and our sample size.)