Review of basic frequentist concepts

Shravan Vasishth

25-29 March 2019

1 Foundations

1.1 Random variable

A random variable X is a function $X : S \rightarrow \mathbb{R}$ that associates to each outcome $\omega \in S$ exactly one number $X(\omega) = x$.

S_X is all the x's (all the possible values of X, the support of X). I.e., $x \in S_X$.
Discrete example: number of coin tosses till H

- $X : \omega \rightarrow x$
- ω: H, TH, TTH,... (infinite)
- $x = 0, 1, 2, \ldots; x \in S_X$

We will write $X(\omega) = x$:
$H \rightarrow 1$

$TH \rightarrow 2$

$\vdots$

The discrete binomial random variable X will be defined by

1. the function $X : S \rightarrow \mathbb{R}$, where S is the set of outcomes (i.e., outcomes are $\omega \in S$).
2. $X(\omega) = x$, and S_X is the **support** of X (i.e., $x \in S_X$).
3. A PMF is defined for X:

$$p_X : S_X \rightarrow [0, 1] \tag{1}$$

$$p_X(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \tag{2}$$

4. A CDF is defined for X:

$$F(a) = \sum_{\text{all } x \leq a} p(x) \tag{3}$$

Continuous example: fixation durations in reading (the normal distribution)

- $X : \omega \to x$
- ω: 145.21, 352.43, 270, ...
- $x = 145.21, 352.43, 270, \ldots; x \in S_X$

The pdf of the normal distribution is:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty \tag{4}$$

We write $X \sim \texttt{norm}(\texttt{mean} = \mu, \texttt{sd} = \sigma)$.

The associated `R` function for the pdf is `dnorm(x, mean = 0, sd = 1)`, and the one for cdf is `pnorm`.

Note the default values for μ and σ are 0 and 1 respectively. Note also that R defines the PDF in terms of μ and σ, not μ and σ^2 (σ^2 is the norm in statistics textbooks).

Table 1: Important R functions relating to random variables.

	Discrete	Continuous
Example:	Binomial(n,θ)	Normal(μ, σ)
Likelihood fn	dbinom	dnorm
Prob X=x	dbinom, pbinom	always 0
Prob $X \geq x, X \leq x, x_1 < X < x_2$	pbinom	pnorm
Inverse cdf	qbinom	qnorm
Generate fake data	rbinom	rnorm

1.2 Maximum likelihood estimate

For the normal distribution, where $X \sim N(\mu, \sigma)$, we can get MLEs of μ and σ by computing:

$$\hat{\mu} = \frac{1}{n} \sum x_i = \bar{x} \tag{5}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 \tag{6}$$

you will sometimes see the "unbiased" estimate (and this is what R computes) but for large sample sizes the difference is not important:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 \tag{7}$$

The significance of these MLEs is that, having assumed a particular underlying pdf, we can estimate the (unknown) parameters (the mean and variance) of the distribution that generated our particular data.

This leads us to the distributional properties of the mean **under repeated sampling**.

1.3 The central limit theorem

For large enough sample sizes, the sampling distribution of the means will be approximately normal, regardless of the underlying distribution (as long as this distribution has a mean and variance defined for it).

1. So, from a sample of size n, and sd σ or an MLE $\hat{\sigma}$, we can compute the standard deviation of the sampling distribution of the means.
2. We will call this standard deviation the estimated **standard error**.
   $SE = \frac{\hat{\sigma}}{\sqrt{n}}$
   I say *estimated* because we are estimating SE using an an estimate of σ.

The standard error allows us to define a so-called **95% confidence interval**:

$$\hat{\mu} \pm 2SE \tag{8}$$

So, for the mean, we define a 95% confidence interval as follows:

$$\hat{\mu} \pm 2\frac{\hat{\sigma}}{\sqrt{n}} \tag{9}$$

What does the 95% CI mean?

2 The t-test

2.1 The hypothesis test

Suppose we have a random sample of size n, and the data come from a $N(\mu, \sigma)$ distribution.

We can estimate sample mean $\bar{x} = \hat{\mu}$ and $\hat{\sigma}$, which in turn allows us to estimate the sampling distribution of the mean under (hypothetical) repeated sampling:

$$N(\bar{x}, \frac{\hat{\sigma}}{\sqrt{n}}) \tag{10}$$

The NHST approach is to set up a null hypothesis that μ has some fixed value. For example:

$$H_0 : \mu = 0 \tag{11}$$

This amounts to assuming that the true distribution of sample means is (approximately) normally distributed and centered around 0, *with the standard error estimated from the data*.

The intuitive idea is that

1. if the sample mean $\bar{x}$ is near the hypothesized μ (here, 0), the data are (possibly) "consistent with" the null hypothesis distribution.
2. if the sample mean $\bar{x}$ is far from the hypothesized μ, the data are inconsistent with the null hypothesis distribution.

We formalize "near" and "far" by determining how many standard errors the sample mean is from the hypothesized mean:

$$t \times SE = \bar{x} - \mu \tag{12}$$

This quantifies the distance of sample mean from μ in SE units.

So, given a sample and null hypothesis mean μ, we can compute the quantity:

$$t = \frac{\bar{x} - \mu}{SE} \tag{13}$$

We will call this the **observed t-value**.

The random varible T:

$$T = \frac{\bar{X} - \mu}{SE} \tag{14}$$

has a t-distribution, which is defined in terms of the sample size n. We will express this as: $T \sim t(n-1)$

Note also that, for large n, $T \sim N(0,1)$.

Thus, given a sample size n, and given our null hypothesis, we can draw t-distribution corresponding to the null hypothesis distribution.

For large n, we could even use N(0,1), although it is traditional in psychology and linguistics to always use the t-distribution no matter how large n is.

2.2 The hypothesis testing procedure

So, the null hypothesis testing procedure is:

1. Define the null hypothesis: for example, $H_0 : \mu = 0$.
2. Given data of size n, estimate $\bar{x}$, standard deviation s, standard error $s/\sqrt{n}$.
3. Compute the t-value:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{15}$$

4. Reject null hypothesis if t-value is large.

2.2.1 Rejection region

So, for large sample sizes, if $\mid t \mid > 2$ (approximately), we can reject the null hypothesis.

For a smaller sample size n, you can compute the exact critical t-value:

```
qt(0.025,df=n-1)
```

This is the critical t-value on the **left**-hand side of the t-distribution. The corresponding value on the right-hand side is:

```
qt(0.975,df=n-1)
```

Their absolute values are of course identical (the distribution is symmetric).

R syntax you should know:

Given iid data y:

```
t.test(y)
```

Given two conditions' paired data vectors cond_a, cond_b (note that the order in which you write the vectors will determine the sign of the observed t-value):

```
t.test(cond_a,cond_b,paired=TRUE)
## identical to above:
t.test(cond_a-cond_b)
```

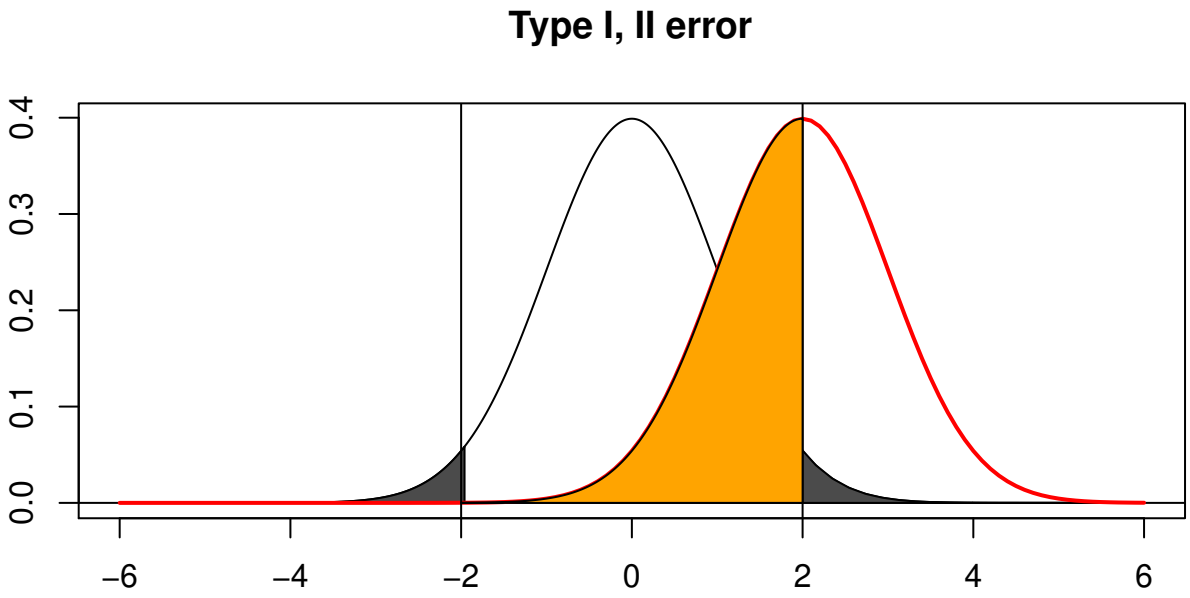
Given data in long form, with the dependent variable written as y and the conditions marked by a column called cond:

```
t.test(y ~ cond,,paired=TRUE)
```

You should know when to aggregate data to meet the one sample (=paired) t-test's assumptions.

3 Type I, II error, power

Reality:	H_0 TRUE	H_0 FALSE
Decision: 'reject':	α **Type I error**	$1 - \beta$ **Power**
Decision: 'fail to reject':	$1 - \alpha$	β **Type II error**

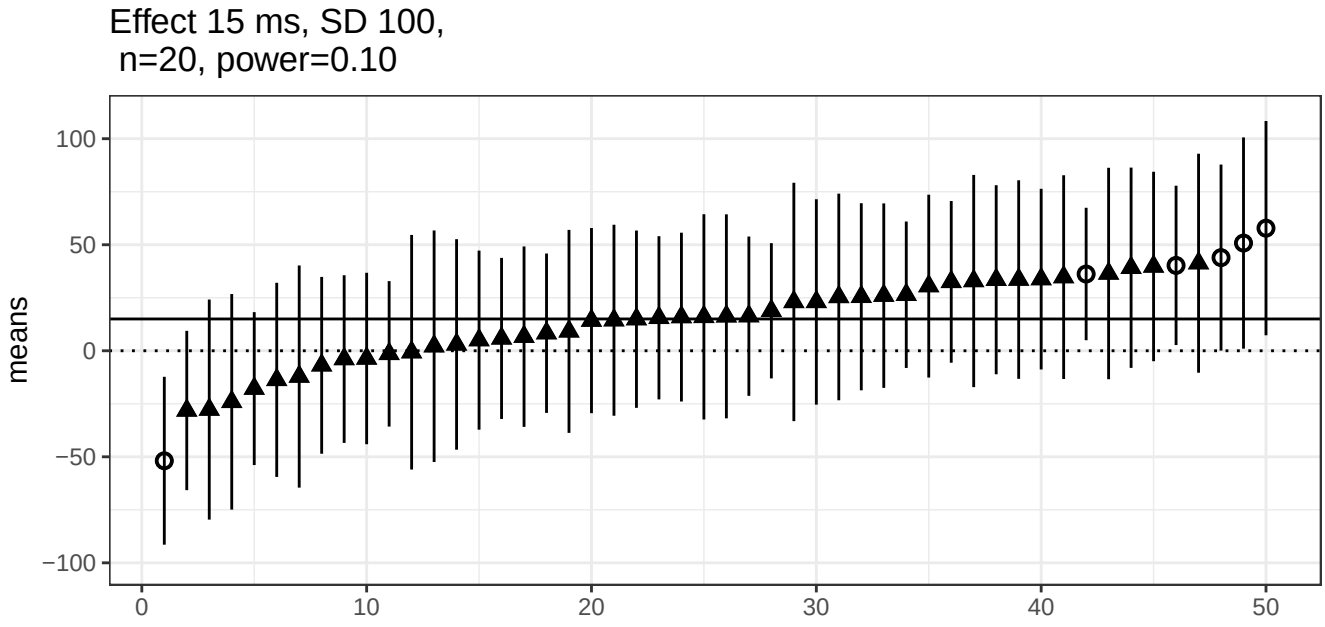


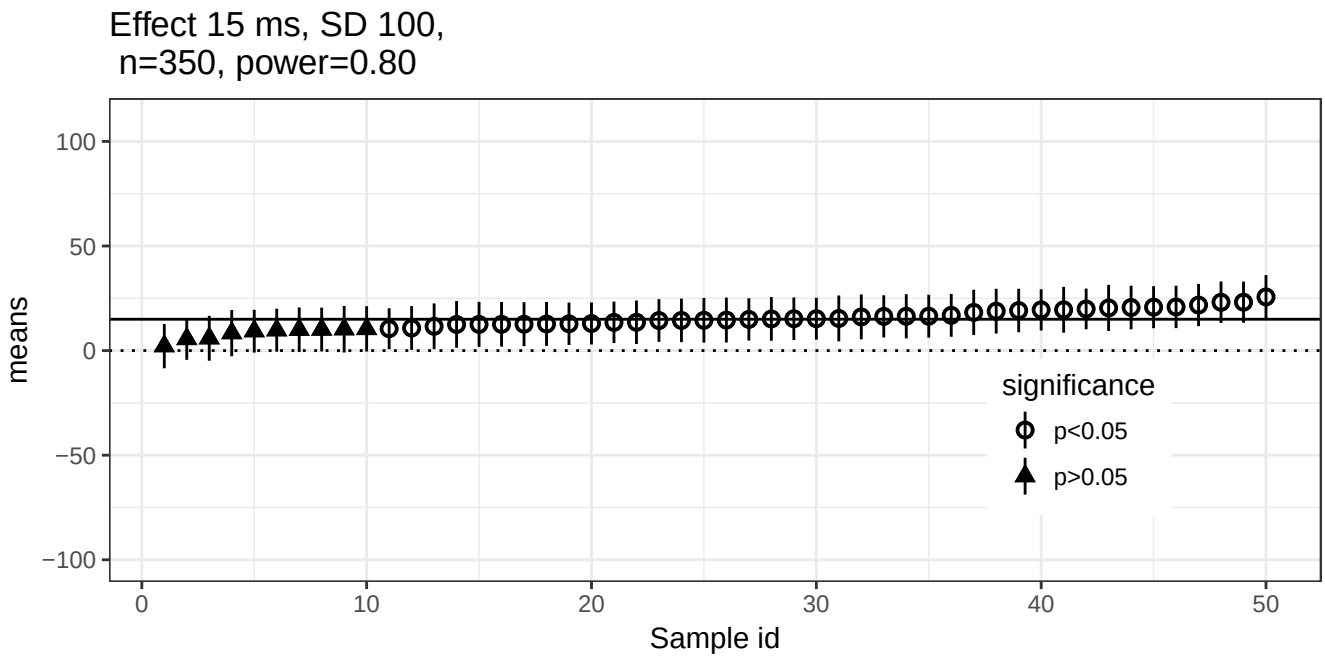
4 Type M, S error

If your true effect size is believed to be D, then we can compute (apart from statistical power) these error rates, which are defined as follows:

1. **Type S error**: the probability that the sign of the effect is incorrect, given that the result is statistically significant.

2. **Type M error**: the expectation of the ratio of the absolute magnitude of the effect to the hypothesized true effect size, given that result is significant. Gelman and Carlin also call this the exaggeration ratio, which is perhaps more descriptive than "Type M error".


Effect 15 ms, SD 100,
n=20, power=0.10
means
100
50
0
−50
−100
0
10
20
30
40
50


Effect 15 ms, SD 100,
n=350, power=0.80
means
100
50
0
−50
−100
0
10
20
30
40
50
Sample id
significance
p<0.05
p>0.05

5 Linear models

We consider the case where we have two conditions (e.g., subject and object relatives), and a repeated measures design.

5.1 Treatment contrast coding

The alphabetically first condition level is coded 0, and the other condition level is coded 1. E.g., if condition labels are obj and subj, then obj is coded 0 and subj 1. You can change this with the command:

```
## code subj as 0 and obj as 1:
dat$condition<-factor(dat$condition,levels=c("subj","obj")
```

In mathematical form, the linear model is:

$$rt = \beta_0 + \beta_1 condition + \epsilon \tag{16}$$

where

- β_0 is the mean for the object relative
- β_1 is the amount by which the object relative mean must be changed to obtain the mean for the subject relative.

The null hypothesis is that the difference in means between the two relative clause types β_1 is:
$H_0 : \beta_1 = 0$
We will make a distinction between the **unknown true mean** β_0, β_1 and the **estimated mean from the data** $\hat{\beta}_0, \hat{\beta}_1$. From the example in the slides:

- Estimated mean object relative processing time: $\hat{\beta}_0 = 471$.
- Estimated mean subject relative processing time: $\hat{\beta}_0 + \hat{\beta}_1 = 471 + -102 = 369$.

5.2 Sum contrast coding

We can code obj as +1 and subj as -1 (or vice versa). Then:

- Estimated **grand mean** processing time: $\hat{\beta}_0 = 420$.
- Estimated mean object relative processing time: $\hat{\beta}_0 + \hat{\beta}_1 = 420 + 51 = 471$.

- Estimated mean subject relative processing time: $\hat{\beta}_0 - \hat{\beta}_1 = 420 - 51 = 369$.

This kind of parameterization is called **sum-to-zero contrast** or more simply **sum contrast** coding. This is the coding we will use.

The null hypothesis for the slope is

$$H_0 : \mathbf{1}\times\mu_{obj} + (-\mathbf{1}\times)\mu_{subj} = 0 \tag{17}$$

The sum contrasts are referring to the ± 1 terms in the null hypothesis:

- object relative: +1
- subject relative: -1

the model is:

Object relative reading times:

$$rt = 420\times\mathbf{1} + 51\times\mathbf{1} + \epsilon \tag{18}$$

Subject relative reading times:

$$rt = 420\times\mathbf{1} + 51\times -\mathbf{1} + \epsilon \tag{19}$$

5.3 Normality assumption of the residuals in the linear models

The model is:

$$rt = \beta_0 + \beta_1 + \epsilon \text{ where } \epsilon \sim Normal(0, \sigma) \tag{20}$$

It is an assumption of the linear model that the residuals are (approximately) normally distributed.

We can check this assumption in R. For a full and formal review of linear models (including linear mixed modeling), see: https://github.com/vasishth/LM