Principe des tests statistiques

Jean Vaillant, Mai 2016

1 Principe des tests statistiques

1.1 Introduction

Un test de signification est une procédure permettant de choisir parmi deux hypothèses celles la plus probable au vu des observations effectuées à partir d'un échantillon ou un dispositif expérimental. Ces deux hypothèses sont disjointes c'est-à-dire s'excluent mutuellement. Les hypothèses auxquelles on s'intéresse portent généralement sur un ou plusieurs paramètres de la population statistique étudiée : ainsi, si l'on s'intéresse à un caractère particulier, on pourra par exemple tester l'égalité de l'espérance du caractère avec une valeur de référence. Par exemple, on peut désirer tester l'égalité d'une contenance attendue de bouteilles commercialisées, avec une valeur de référence en l'occurrence la contenance indiquée sur l'étiquette commerciale. Un inspecteur de la direction de la consommation peut choisir un certain nombre de bouteilles dans la production de l'usine concernée. Sachant qu'il y a un aléa d'échantillonnage et une variabilité dans le système de remplissage des bouteilles, comment tranchera t-il entre l'hypothèse *la contenance attendue est égale à la contenance annoncée* et l'alternative contraire?

1.2 Erreurs décisionnelles et risques

Le principe de base d'un test de signification est de considérer une hypothèse privilégiée H_0 et une alternative H_1, puis de bâtir une règle permettant de décider de rejeter ou pas H_0. Le tableau 1 résume les 4 situations possibles. L'erreur de première espèce est de rejeter l'hypothèse privilégiée H_0 alors qu'elle est vraie. L'erreur de seconde espèce est de ne pas rejeter H_0 alors qu'elle est fausse. α est la probabilité de rejeter à tort l'hypothèse H_0; α est aussi appelé risque de première espèce, ou niveau du test. β est la probabilité de ne pas rejeter H_0 alors que l'hypothèse alternative H_1 est vraie; β est appelé risque de seconde espèce. La valeur $1-\beta$ est la puissance du test, et traduit la faculté de rejeter H_0 quand l'alternative H_1 est vraie.

Dans la pratique, α est fixé par l'expérimentateur (les valeurs les plus courantes sont 0,05 ou 0,01. On dit qu'on contrôle le risque de première espèce. Par contre, β peut être difficile à calculer. Heureusement, ce calcul n'est pas nécessaire sauf si l'on veut comparer plusieurs procédure de tests.

Dans la littérature, H_0 est aussi appelée *hypothèse nulle* ou encore *hypothèse principale*. Elle joue un rôle prédominant par rapport à l'hypothèse H_1 qui est souvent l'hypothèse alternative contraire. On cherche à contrôler le risque α de rejeter à tort H_0 en lui imposant une valeur relativement faible (au plus 0,05). Le fait d'imposer une valeur faible à α conduit à n'abandonner l'hypothèse H_0 que dans des cas qui *semblent sortir nettement de l'ordinaire* si H_0 était vraie.

		Etat de la nature	
		H_0	H_1
Décision	Rejet de H_0	α	$1-\beta$
	Non rejet de H_0	$1-\alpha$	β

Table 1: Risques décisionnels conditionnels à l'état de la nature

1.3 Probabilité critique (ou p-value ou niveau de signification observé)

Notons bien que plus α est choisi petit, plus la règle de décision est stricte (ou conservative) dans la mesure où elle aboutit à rejeter H_0 que dans des cas rarissimes et donc à conserver cette hypothèse quelque fois à tort. Une vision moderne, liée à l'explosion de la puisssance des ordinateurs et de processus numériques d'approximation rapides et précis, est d'afficher la p-value ou probabilité critique p_c. Par définition, **la p-value est la plus petite des valeurs de risque de première espèce pour lesquelles la décision serait de rejeter** H_0. La valeur p_c est calculée à partir des observations et de leurs propriétés distributionnelles sous H_0. Comme p_c est le plus petit niveau de signification auquel on rejette l'hypothèse H_0, il est aussi appelé *niveau de signification observé*. L'amélioration fulgurante des capacités de calcul permet maintenant de baser les règles de décision sur les probabilités critiques sans forcément comparer la statistique de test avec une valeur seuil, comme cela se faisait classiquement.

La définition formelle de la p-value donnée ci-dessus est difficile à ingurgiter et peut conduire à une mauvaise utilisation et/ou une mauvaise interprétation de l'inférence statistique ([3]). Une définition littérale et plus parlante aux non initiés peut être la suivante : **la p-value est une mesure de la compatibilité des données avec l'hypothèse privilégiée**. Plus cette p-value est proche de zéro, plus la compatibilité est faible et donc conduit à rejeter cette hypothèse. La proximité à zéro dépend de la sévérité que l'on s'impose à travers le risque α.

1.4 Critère de test, Région critique

Tout test d'une hypothèse H_0 est basé sur un critère C qui est calculé à partir des observations effectuées. C est appelé critère de test (ou statistique de test). C est une quantité dépendant des données observées ou recueillies lors de l'expérimentation ou l'enquête. C'est donc une variable aléatoire dont la valeur observée nous permettra de déterminer quelle hypothèse est la plus plausible, en se référant à la distribution de probabilité de cette variable aléatoire sous H_0. La prise de décision se fera selon une règle dont la forme est généralement :

$$\begin{cases} \text{Rejet de } H_0 & \text{si } C \in R_c(\alpha) \\ \text{Non Rejet de } H_0 & \text{si } C \notin R_c(\alpha) \end{cases}$$

où $R_c(\alpha)$ est donc l'ensemble des valeurs pour lesquelles la statistique de test conduit au rejet de l'hypothèse H_0 au niveau de signification α. Cet ensemble $R_c(\alpha)$ est donc appelé région critique (ou zone de rejet) du test au niveau α.

Le complémentaire de $R_c(\alpha)$ est l'ensemble des valeurs pour lesquelles la statistique de test conduit au non rejet de l'hypothèse H_0. On l'appelle région (ou zone) d'acceptation du test au niveau α.

La région critique ou zone de rejet correspond donc aux valeurs de C qui seraient trop extraordinaires sous l'hypothèse H_0 pour être considérées comme le fruit du hasard d'échantillonnage.

Notons que les logiciels statistiques modernes calculent la p-value p_c et fournissent la règle de décision de niveau α sous la forme :

$$\begin{cases} \text{Rejet de } H_0 & \text{si } p_c < \alpha \\ \text{Non Rejet de } H_0 & \text{si } p_c \geq \alpha \end{cases}$$

1.5 Test unilatéral, test bilatéral

Rappelons que la région de rejet $R_c(\alpha)$ d'un test de niveau α basé sur la statistique de T est l'ensemble des valeurs possibles de T pour lesquelles la règle de décision nous conduit à rejeter H_0 au niveau α.

Un test est dit unilatéral si cette région de rejet $R_c(\alpha)$ est entièrement située à une des extrémités de la distribution d'échantillonnage de T.

Un test est dit bilatéral si cette région est située aux deux extrémités de la distribution d'échantillonnage de T.

La figure 1 indique des régions critiques de niveau 5% basées sur un critère de test suivant la loi normale centrée réduite sous H_0.

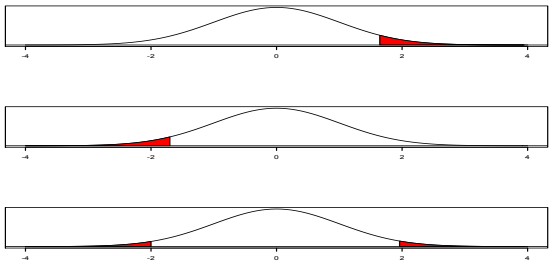

Figure 1: Régions critiques (en rouge) pour un critère de loi N(0,1) et $\alpha = 0,05$.

1.6 Liens entre p-value et statistique de test dans quelques cas simples

Soit H_0 une hypothèse privilégiée. On désire tester cette hypothèse vis à vis d'une hypothèse alternative. Soit T la statistique de test (à valeurs réelles) utilisée pour effectuer le test de H_0, c'est-à-dire vérifier si H_0 est vraie. On désigne par P_0 la loi de probabilité de T sous H_0.

A partir des données recueillies, on a une valeur observée t pour la statistique de test. Le principe général des tests d'hypothèse est de rejeter l'hypothèse H_0 quand t est en extrémité de distribution de P_0 et correspond donc à une valeur fort peu probable sous H_0. Pour quantifier "les chances d'occurrence" d'une telle valeur t sous H_0, on calcule la probabilité critique p_c dont la définition suit :

On a vu que la probabilité critique p_c (ou p-value) d'un test d'hypothèse pour des observations données est le plus petit des niveaux de signification pour lesquels la décision est de rejeter H_0. Autrement dit, p_c est la plus petite probabilité, au vu des observations, de rejeter à tort l'hypothèse privilégiée H_0. Le lien avec la loi de la statistique de test T et la statistique observée est le suivant : la p-value est la probabilité qu'une réalisation de la statistique de test T soit plus *extraordinaire* (c'est-à-dire plus en extrémité de distribution) que la valeur observée t sous l'hypothèse H_0.

Ainsi, de très faibles valeurs pour p_c indiquent que l'hypothèse privilégiée H_0 est peu probable. Plus p_c est faible, plus les données témoignent que le phénomène observé a très peu de chances de se produire sous H_0. Elles nous conduisent alors à rejeter H_0.

Pour définir de façon rigoureuse le lien entre probabilité critique p_c et statistique de test, il est nécessaire, et c'est le cas pour toute expérience aléatoire ξ, d'introduire l'espace probabilisé $(\Omega, \mathcal{A}, P_0)$ où Ω est l'ensemble des résultats possibles de ξ, où $\mathcal{A}$ est la tribu d' événements associés à Ω, et P_0 la loi de probabilité sous H_0. Le résultat observé de

l'expérience ξ est noté ω^* et on a donc $t = T(\omega^*)$.

On a indiqué précédemment qu'un test basé sur la statistique T est dit unilatéral si, pour tout niveau α, sa région de rejet est entièrement située à une des extrémités de la distribution de probabilité de T. Il est dit bilatéral si cette région de rejet est située aux deux extrémités de la distribution de probabilité de T..

Considérons les trois cas simple suivants :

1) Pour un test unilatéral droit,
$$p_c = P_0(\{\omega \in \Omega \quad | \quad T(\omega) > t\}).$$
2) Pour un test unilatéral gauche,
$$p_c = P_0(\{\omega \in \Omega \quad | \quad T(\omega) < t\}).$$
3) Pour un test bilatéral, avec T de loi centrée symétrique sous H_0,
$$p_c = P_0(\{\omega \in \Omega \quad | \quad |T(\omega)| > |t|\}).$$

Rappelons que α le niveau de signification du test est par définition la probabilité de rejeter H_0 alors que H_0 est vraie. Si la probabilité critique p_c est plus petite que le niveau de signification α, alors l'hypothèse H_0 est rejetée.

Notons F_0 la fonction de répartition de T sous H_0, et examinons le lien entre région critique d'un test, règle de décision et probabilité critique.

1) **Test unilatéral droit**

   La région critique du test est de la forme $]c_{1-\alpha}, +\infty[$ avec $c_{1-\alpha}$ fractile d'ordre $1-\alpha$ de la loi F_0 c'est-à-dire $F_0(c_{1-\alpha}) = 1-\alpha$.

   La probabilité critique p_c du test unilatéral droit est par définition :

   $$p_c = P_0(T > t) = 1 - F_0(t).$$

   On démontre aisément le résultat suivant :

   $$p_c < \alpha \quad \Rightarrow \quad t > c_{1-\alpha} \qquad (1)$$

   Preuve :

   $p_c < \alpha \Leftrightarrow 1-F_0(t) < \alpha \Leftrightarrow 1-\alpha < F_0(t) \Leftrightarrow F_0(c_{1-\alpha}) < F_0(t) \Rightarrow c_{1-\alpha} < t.$

2) **Test unilatéral gauche**

La région critique du test est $]-\infty, c_\alpha[$ avec $F_0(c_\alpha^-) = \alpha$.

La probabilité critique p_c du test est par définition :

$$p_c = P_0(T < t) = F_0(t^-).$$

On démontre que :

$$p_c < \alpha \quad \Rightarrow \quad t < c_\alpha \tag{2}$$

Preuve :

$$p_c < \alpha \Leftrightarrow F_0(t^-) < \alpha \Leftrightarrow F_0(t^-) < F_0(c_\alpha^-) \Rightarrow t < c_\alpha.$$

3) **Test bilatéral avec T de loi symétrique centrée sous H_0**

La région critique du test est $]-\infty, -c_{1-\alpha/2}[\cup]c_{1-\alpha/2}, +\infty[$

où $c_{1-\alpha/2}$ est le fractile d'ordre $1-\alpha/2$ de F_0 c'est-à-dire

$F_0(c_{1-\alpha/2}) = 1 - \dfrac{\alpha}{2}$. La probabilité critique p_c est par définition :

$$p_c = P_0(|T| > |t|) = P_0(T > |t|) + P_0(T < -|t|) = 1 - F_0(|t|) + F_0(-|t|^-).$$

On a le résultat suivant :

$$p_c < \alpha \Rightarrow |t| > c_{1-\alpha/2}. \tag{3}$$

Preuve :

$p_c < \alpha \Leftrightarrow 1 - F_0(|t|) + F_0(-|t|^-) < \alpha \Leftrightarrow 1 - F_0(|t|) + 1 - F_0(|t|^+) < \alpha$

or F_0 est continue à droite (en tant que fonction de répartition) d'où

$2(1-F_0(|t|)) < \alpha \Leftrightarrow 1 - \dfrac{\alpha}{2} < F_0(|t|) \Leftrightarrow F_0(c_{1-\alpha/2}) < F_0(|t|) \Rightarrow c_{1-\alpha/2} < |t|.$

Les résultats (1), (2) et (3) confirment que, pour un test de niveau α, l'hypothèse H_0 est rejetée dès lors que la probabilité critique est inférieure strictement à α. Ceci est très utile dans la pratique : la règle de décision consiste simplement à comparer p_c à α au lieu de comparer t à des valeurs seuils fournies par des tables de fractiles de lois usuelles. Les logiciels statistiques calculent et présentent donc ces probabilités critiques, qui sont difficiles à obtenir sans moyen de calcul approprié.

Une autre utilisation de la probabilité critique en théorie de la décision consiste non plus à la comparer avec un seuil de signification mais de la combiner, en tant qu'indice-témoin, avec d'autres sources d'information.

1.7 Exemple du dé supposé pipé sur le 1

On désire tester si un dé cubique numéroté de 1 à 6 est pipé sur le 1, en privilégiant l'hypothèse selon laquelle il est équilibré. En notant p_1 la probabilité d'apparition du 1 lors d'un lancer, le problème de test est donc :

$H_0 : p_1 = 1/6$ contre $H_1 : p_1 \neq 1/6$.

Si on lance 20 fois le dé et que l'on considère la statistique de test T égale au nombre de 1 obtenus, alors sous H_0 le critère T suit la loi binomiale de paramètres 20 et 1/6.

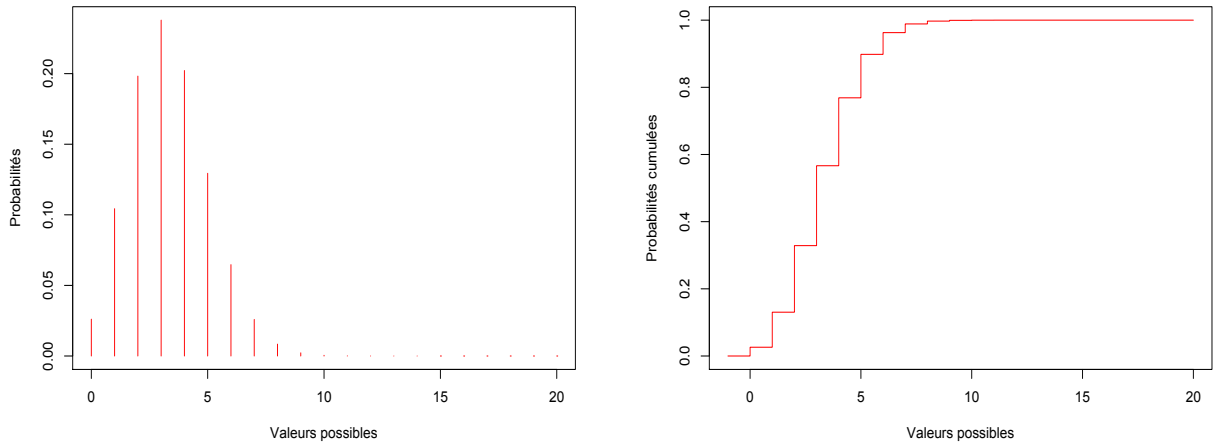

Figure 2: Représentations de la loi binomiale de paramètres 20 et 1/6.

La région critique de niveau $0,02$ est $R_c(0,02) = [\![8;20]\!]$. Si l'on réduit le niveau à 0,001, on a alors $R_c(0,001) = [\![10;20]\!]$ Les valeurs 8 et 9 pour T qui étaient extraordinaires au niveau $0,02$, ne le sont plus au niveau $0,001$.

1.8 Tests de comparaison d'une proportion à une valeur de référence

On considère une population statistique pour laquelle une proportion inconnue p d'individus vérifie une certaine propriété (par exemple sont atteints d'une maladie ou sont favorables à un projet). On désire comparer la valeur inconnue p à une valeur de référence p_0. Par exemple, en épidémiologie, p est la prévalence d'une maladie et p_0 un seuil d'alerte sanitaire.

On choisit dans cette population un nombre n d'individus par tirages indépendants. On note X_n le nombre d'individus dans cet échantillon vérifiant la propriété étudiée. X_n suit la loi binomiale de paramètres n et p. On peut s'intéresser aux trois problèmes de test suivants :

Problème 1. $H_0 : p = p_0$ contre $H_1 : p \neq p_0$.

Problème 2. $H_0 : p \leq p_0$ contre $H_1 : p > p_0$.

Problème 3. $H_0 : p \geq p_0$ contre $H_1 : p < p_0$.

Ainsi, le problème de dé truqué vu au paragraphe précédent correspond au problème 1 avec $p_0 = 1/6$. Pour les trois problèmes de test, X_n est un critère de test pertinent pour H_0. Comme on peut approcher la loi binomiale par la loi normale de même espérance et variance pour n suffisamment grand (théorème de De Moivre-Laplace, [1]), alors quand tel est le cas, la variable aléatoire associée standardisée $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$ suit approximativement la loi $\mathcal{N}(0,1)$. Pour $p = p_0$, on peut écrire

$$Z_n = \frac{\widehat{p}_n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}, \tag{4}$$

où $\widehat{p}_n$ est la proportion observée dans l'échantillon de taille n.

L'expression (4) fait clairement apparaître Z_n comme un écart pondéré entre la proportion observée $\widehat{p}_n$ et la proportion de référence p_0. Pour chacun des trois probèmes de test, on peut donc utiliser Z_n comme critère de test, ce qui nous donne (en notant u_p le fractile d'ordre p de la loi $\mathcal{N}(0,1)$ et Φ sa fonction de répartition) les règles de décision de niveau α suivantes.

Pour le Problème 1,

$$\begin{cases} \text{Rejet de } H_0 & \text{si } |Z_n| > u_{1-\alpha/2} \\ \text{Non Rejet de } H_0 & \text{si } |Z_n| \leq u_{1-\alpha/2}. \end{cases} \tag{5}$$

La p-value est $p_c = 2\big(1 - \Phi(|Z_n|)\big)$.

Pour le Problème 2,

$$\begin{cases} \text{Rejet de } H_0 & \text{si } Z_n > u_{1-\alpha} \\ \text{Non Rejet de } H_0 & \text{si } Z_n \leq u_{1-\alpha}. \end{cases} \tag{6}$$

La p-value est $p_c = 1 - \Phi(Z_n)$.

Pour le Problème 3,

$$\begin{cases} \text{Rejet de } H_0 & \text{si } Z_n < -u_{1-\alpha} \\ \text{Non Rejet de } H_0 & \text{si } Z_n \geq -u_{1-\alpha}. \end{cases} \tag{7}$$

La p-value est $p_c = \Phi(Z_n)$.

Application : Dans une parcelle forestière, des arbres d'une certaine espèce sont attaqués, en proportion inconnue, par un parasite. On redoute que le seuil de nuisibilité de 20% ne soit dépassé car un traitement phytosanitaire coûteux devra alors être appliqué. Mais on privilégie l'hypothèse selon laquelle le seuil de nuisibilité n'est pas atteint. On choisit indépendamment 50 arbres de cette espèce et on en trouve 14 qui sont attaqués par le parasite. On se fixe un risque de première espèce de 5%. Quelle décision doit on prendre?

Nous faisons face au problème 2 avec $p_0 = 0,2$ et $\alpha = 0,05$. La taille n de l'échantillon vaut 50 et la proportion observée $\widehat{p}_n$ est égale à $14/50 = 0,28$. La statistique de test est $Z_{50} = \dfrac{0,28 - 0,20}{\sqrt{0,20(1-0,20)}}\sqrt{50} = 1,414$ et $u_{1-0,05} = u_{0,95} = 1,645$ donc on ne rejette pas l'hypothèse $p \leq 0,2$ au niveau $0,05$.

La proportion observée 28% n'est pas significativement supérieure à 20% au niveau 5%. On décide de ne pas effectuer de traitement phytosanitaire.

Remarquons que, d'un point de vue pratique, il est plus simple de prendre une décision en se basant sur la valeur observée X_{50} pour le nombre d'arbres parasités dans l'échantillon plutôt que de calculer Z_{50}. En tenant compte des expressions (4) et (6), pour le problème 2, on rejette H_0 au niveau α si

$$\frac{X_n/n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} > u_{1-\alpha} \quad \text{c'est-à-dire si} \quad X_n > np_0 + u_{1-\alpha}\sqrt{np_0(1-p_0)}.$$

La région critique de niveau α du test, pour le problème 2 et le critère de test X_n est donc

$$R_c(\alpha) =]np_0 + u_{1-\alpha}\sqrt{np_0(1-p_0)}; +\infty[.$$

Dans le cas présent, on a donc

$$R_c(0,05) =]50 \times 0,2 + 1,645 \times \sqrt{50 \times 0,2 \times (1-0,2)}; +\infty[=]14,65; +\infty[.$$

On décidera d'effectuer un traitement phytosanitaire au niveau 5% si $X_{50} \geq 15$. L'ensemble des valeurs critiques au niveau 5% pour le nombre d'arbres parasités est $[\![15; 50]\!]$.

1.9 Outils avec R

Dans l'environnement de calcul et programmation R (https://cran.r-project.org, [2]), on peut écrire sa propre fonction pour exécuter un test statistique en faisant appel à des fonctions dites natives sous R (fonctions prédéfinies prêtes à l'emploi). Pour ce qui concerne les tests classiques d'hypothèses, il existe de nombreuses fonctions dans le package de base, sans compter celles pouvant être disponibles en installant des packages spécifiques. Citons quelques unes des fonctions du package "stats" par ordre alphabétique :

1. chisq.test(), test du χ^2 d'ajustement, test du χ^2 d'indépendance,
2. cor.test(), test de corrélation pour échantillons appariés,
3. kruskal.test(), test des rangs de Kruskal-Wallis pour la comparaison de distributions,
4. ks.test(), test d'ajustement de Kolmogorov-Smirnov,
5. prop.test(), test de comparaison de proportions,
6. shapiro.test(), test de normalité de Shapiro-Wilk,
7. t.test(), test t de Student pour la comparaison de moyennes,
8. var.test(), test de Fisher de comparaison de variances pour deux variables gaussiennes,
9. wilcox.test(), test de rangs de Wilcoxon, tests de Mann-Whitney.

Notons que, sous R, on peut par de simples requêtes calculer la p-value pour certaines procédures de test statistique. Si on revient au problème de franchissement de seuil phytosanitaire vu précédemment ($p \leq 0,2$ contre $p > 0,2$), on peut calculer la p-value en tapant la requête suivante :

```
1-pnorm((14-50*0.2)/sqrt(50*0.2*(1-0.2)))
```

où "pnorm" est la fonction native calculant la valeur de la fonction de répartition de la loi normale. On obtient la valeur 0,079 qui est supérieure à 0,05 donc on ne rejette pas l'hypothèse $p \leq 0,2$.

References

[1] Gilbert Saporta, *Probabilités, analyse des données et statistique*, Technip, Paris, France, 3^{rd} edition ed., 2011.

[2] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015.

[3] Ronald L. Wasserstein and Nicole A. Lazar, *The ASA's statement on p-values: context, process, and purpose*, The American Statistician, (2016).