# The Measurement of Statistical Evidence
# Lecture 4 - part 1

Michael Evans
University of Toronto
http://www.utstat.utoronto.ca/mikevans/sta4522/STA4522.html

2021

## 3. p-values and confidence

- the p-value is currently a basic tool of inference and yet serious reservations have been raised to the extent that one journal banned its use as a "measure of evidence" due to the replicability crisis

- the current approach taken by the statistical profession is to suggest that there is nothing wrong with p-values rather it is the users who do not understand how to use them correctly

- so what is a p-value?

**Definition** Suppose there is a hypothesis $H_0 \subset \Theta$ concerning the true value of $\theta$ for the model $\{f_\theta : \theta \in \Theta\}$ and a statistic $T$ whose probability distribution $P_{H_0}$ is known and fixed for each $\theta \in H_0$ and such that extreme values correspond to large values of $T$. Then $H_0$ is assessed by computing the *p-value* $P_{H_0}(T \geq T(x))$ for observed value $T(x)$.

- if $P_{H_0}(T \geq T(x))$ is small, then it is concluded that there is evidence against $H_0$

**Question 1**: How small is small enough?

- a *rejection trial* adds the ingredient of a value $\alpha \in [0, 1]$ s.t. if $P_{H_0}(T \geq T(x)) \leq \alpha$, then evidence against is concluded

- historically $\alpha = 0.05$ has been used but a recent recommendation has been that this be replaced by $\alpha = 0.005$

- will this work?

**Example** *Cornfield (1966)*

- $x = (x_1, \ldots, x_n) \overset{i.i.d.}{\sim} N(\mu, \sigma_0^2)$ with $\mu \in R^1, \sigma_0^2$ known and $H_0 = \{\mu_0\}$
- then with $T_n(x) = \sqrt{n}|\bar{x} - \mu_0|/\sigma_0 \sim |Z|$ where $Z \sim N(0, 1)$ the p-value is the $Z$-test

$$P_{H_0}(T_n \geq T_n(x)) = P(|Z| \geq \sqrt{n}|\bar{x} - \mu_0|/\sigma_0) = 2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_0|/\sigma_0))$$

which $\leq 0.05$ when $\sqrt{n}|\bar{x} - \mu_0|/\sigma_0 \geq z_{0.975}$

- suppose an investigator collects $n$ data values, performs the $Z$-test and gets a p-value of 0.06

- this is close to the 0.05 level so they decide to collect $m$ additional data values and compute a new $Z$-test based on the $n + m$ values obtaining p-value $< 0.05$ and the result is submitted for publication

- but this is a two-stage test and, when $H_0$ is true, the probability of evidence against $\mu_0$ is

$$\begin{aligned} & P_{H_0}(T_n \geq z_{0.975}) + P_{H_0}(T_{m+n} \geq z_{0.975} \mid T_n < z_{0.975}) P_{H_0}(T_n < z_{0.975}) \\ = \; & 0.05 + P_{H_0}(T_{m+n} \geq z_{0.975} \mid T_n < z_{0.975})(0.95) > 0.05 \end{aligned}$$

and so evidence against $H_0$ can never be found at the 0.05 level

- the problem here is the use of the 5% level to determine evidence against and this problem persists no matter what $\alpha$ level is used, yet collecting additional data in such circumstances seems like a very natural thing to do

**Question 2**: Why isn't a large p-value $(\geq \alpha)$ evidence in favor?

- suppose the probability measure $P_{H_0}$ for $T$ is continuous with cdf $F_{H_0}$

- then $P_{H_0}(T \geq T(x)) = 1 - F_{H_0}(T(x))$ so when $H_0$ is true the probability distribution of the p-value is when $\theta \in H_0$

$$P_\theta(1 - F_{H_0}(T(X)) \leq u) = P_{H_0}(F_{H_0}(T) \geq 1 - u) = u$$

since $F_{H_0}(T) \sim U(0,1)$ when $H_0$ is true

- so when $H_0$ is true all possible values of the p-value are equally likely, independent of the amount of data while, when when $H_0$ is false, the p-value typically converges to 0 as the amount of data increases

**Question 3**: Do p-values measure scientific significance or just statistical significance?

- suppose in the Z-test $\mu_{true} = \mu_0 + \delta$ and $\delta$ is very small, then for $n$ large enough $P_{H_0}(T_n \geq T_n(x)) < \alpha$ even when the difference $\delta$ is scientifically irrelevant

- so p-values measure statistical significance not scientific significance

*Boring, E. (1919) Mathematical vs statistical significance. Psychological Bulletin, 16, 10, 335-338.*

- the common recommendation to deal with this issue is to compute a confidence interval for the parameter of interest but this doesn't really help unless you know the difference that matters $\delta$ and even then it is ambiguous as some values in the CI may be relevant and some not

- the real solution is to incorporate $\delta$ into the measure of evidence, for example, put $H_0 = [\mu_0 - \delta, \mu_0 + \delta]$ and assess the evidence in favor or against, but this isn't done with p-values

- basic to resolving all these issues is to use a valid measure of evidence which the p-value isn't

**Definition** A map $C : \mathcal{X} \rightarrow 2^{\Psi}$ is a *$\gamma$-confidence region* for $\psi = \Psi(\theta)$ if $P_{\theta}(\Psi(\theta) \in C(X)) \geq \gamma$ for every $\theta \in \Theta$.

- when $x$ is observed then record $C(x)$ as "typically" the estimate is in $C(x)$ and so the "size" of $C(x)$ serves as a measure of the accuracy of the estimate

**Example** *Absurd confidence intervals*

- the model $\mathcal{X} = R^1, f_{\theta}(x) = (1-\theta)f(x) + \theta f(x-1)$ where $f$ is the $N(0,1)$ density function and $\Theta = [0,1]$
- Plante(1991) a 0.95-confidence interval for $\theta$ that is uniformly most accurate and unbiased is given by

$$C(x) = \begin{cases} [0,1] & -1.68148 \leq x \leq 2.68148 \\ \phi & \text{otherwise} \end{cases}$$

**Example** *Fieller (1954) Some problems in interval estimation. JRSSB, 16, 2, 175–185.*

- $x = (x_1, \ldots, x_m) \overset{i.i.d.}{\sim} N(\mu, \sigma_0^2)$ ind. of $y = (y_1, \ldots, y_n) \overset{i.i.d.}{\sim} N(\nu, \sigma_0^2)$ and $\psi = \Psi(\mu, \nu) = \mu/\nu$ various frequentist approaches produce absurd confidence intervals (sometimes equal to $R^1$)

## 4. Bayesian Inference

- the prior $\pi$ (a proper probability distribution on $\Theta$) is added to the ingredients, model $\{f_\theta : \theta \in \Theta\}$ and data $x$

- gives a joint prior probability distribution $(\theta, x) \sim \pi(\theta) f_\theta(x)$

- recall the prior $\pi$ expresses our beliefs about the true value of $\theta$

- after observing $x$ the principle of conditional probability implies we replace $\pi$ by the posterior

$$\pi(\theta \,|\, x) = \frac{\pi(\theta) f_\theta(x)}{m(x)}$$

where $m(x) = \int_\Theta \pi(\theta) f_\theta(x)\, d\theta$ is the prior predictive distribution of $x$

- how to choose a prior? elicitation

**Example** *location normal*

- $x = (x_1, \ldots, x_n) \overset{i.i.d.}{\sim} N(\mu, \sigma_0^2)$ with $\mu \in R^1, \sigma_0^2$ known and $\pi$ a $N(\mu_0, \tau_0^2)$ dist. so

$$\pi(\mu \,|\, x) \propto \pi(\mu) f_\mu(x) \propto \exp\left\{-\frac{(\mu - \mu_0)^2}{2\tau_0^2}\right\} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

and using $\sum_{i=1}^{n}(x_i-\mu)^2 = n(\bar{x}-\mu)^2 + \sum_{i=1}^{n}(x_i-\bar{x})^2$

$$\pi(\mu \,|\, x) \propto \exp\left\{-\frac{1}{2}\left[\frac{(\mu-\mu_0)^2}{\tau_0^2} + \frac{n(\bar{x}-\mu)^2}{\sigma_0^2}\right]\right\}$$

and

$$\begin{aligned} & \frac{(\mu-\mu_0)^2}{\tau_0^2} + \frac{n(\bar{x}-\mu)^2}{\sigma_0^2} \\ = \;& \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma_0^2}\right)\mu + \left(\frac{\mu_0^2}{\tau_0^2}+\frac{(n\bar{x})^2}{\sigma_0^2}\right) \\ = \;& \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)\left(\mu - \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma_0^2}\right)\right) + \text{constant} \end{aligned}$$

and so putting

$$\mu_x = \tau_x^2\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma_0^2}\right), \tau_x^2 = \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}$$

then $\mu \,|\, x \sim N\left(\mu_x, \tau_x^2\right)$

- how to choose the hyperparameters $(\mu_0, \tau_0^2)$?

- recall the data is the result of a measurement process so an observation will fall in some known interval $(l, u)$ with "virtual certainty" (prob. 0.99)

- so one possibility is $\mu_0 = (l+u)/2$ and choose $\tau_0$ so that $\Phi((u-\mu_0)/\tau_0) - \Phi((l-\mu_0)/\tau_0) = 0.99$ (conservative)

- e.g. $(l, u) = (3, 10)$ so $\mu_0 = 6.5$ and

$$\begin{aligned}
0.99 &= \Phi((10-6.5)/\tau_0) - \Phi((3-6.5)/\tau_0) \\
&= \Phi(3.5)/\tau_0) - \Phi(-3.5)/\tau_0) = 2\Phi(3.5)/\tau_0) - 1 \\
\tau_0 &= 3.5/\Phi^{-1}(0.995) = 1.358786
\end{aligned}$$

- so the $N(6.5, 1.35878^2)$ expresses prior beliefs about $\mu$

- if $\sigma_0^2 = 2, n = 10, \bar{x} = 7.3$ is observed, then the posterior of $\mu$ is $N\left(\mu_x, \tau_x^2\right) = N(7.23, 0.18)$

- for a marginal parameter $\psi = \Psi(\theta)$ we have the marginal prior and posterior

$$\begin{aligned} \pi_\Psi(\psi) &= \int_{\{\theta:\psi=\Psi(\theta)\}} \pi(\theta) J_\Psi(\theta)\, d\theta \\ \pi_\Psi(\psi \,|\, x) &= \int_{\{\theta:\psi=\Psi(\theta)\}} \pi(\theta \,|\, x) J_\Psi(\theta)\, d\theta \end{aligned}$$

where $J_\Psi(\theta)$ is a volume distortion factor (see text Appendix)

- two properties

(1) Consistency: the posterior for $\psi$ is the same as if we start with the ingredients $(\{m(\cdot \,|\, \psi) : \psi \in \Psi\}, \pi_\Psi, x)$ where

$$\begin{aligned} m(x \,|\, \psi) &= \int_{\{\theta:\psi=\Psi(\theta)\}} \pi(\theta \,|\, \psi) f_\theta(x)\, d\theta \\ \pi(\theta \,|\, \psi) &= \frac{\pi(\theta) J_\Psi(\theta)}{\pi_\Psi(\psi)} \end{aligned}$$

(the "nuisance" parameters have been integrated out)
Proof:

$$\begin{aligned} \pi_\Psi(\psi \,|\, x) &= \int_{\{\theta:\psi=\Psi(\theta)\}} \pi(\theta \,|\, x) J_\Psi(\theta)\, d\theta \\ &= \int_{\{\theta:\psi=\Psi(\theta)\}} \frac{\pi(\theta) f_\theta(x)}{m(x)} J_\Psi(\theta)\, d\theta \\ &= \frac{\pi_\Psi(\psi)}{m(x)} \int_{\{\theta:\psi=\Psi(\theta)\}} \frac{\pi(\theta) J_\Psi(\theta)}{\pi_\Psi(\psi)} f_\theta(x)\, d\theta \\ &= \frac{\pi_\Psi(\psi) m(x \,|\, \psi)}{m(x)} \end{aligned}$$

(2) if after observing $x$, new independent data $y$ is observed with model $\{g_\theta : \theta \in \Theta\}$, then the posterior for $\psi$ based on $(x, y)$ is

$$
\begin{aligned}
\pi_\Psi(\psi \,|\, x, y) &= \frac{\pi_\Psi(\psi) m(x, y \,|\, \psi)}{m(x, y)} = \frac{\pi_\Psi(\psi \,|\, x) m(x)}{m(x \,|\, \psi)} \frac{m(x, y \,|\, \psi)}{m(x, y)} \\
&= \frac{\pi_\Psi(\psi \,|\, x) m(y \,|\, \psi, x)}{m(y \,|\, x)}
\end{aligned}
$$

(so the posterior for $\psi$ based on $x$ now serves as a prior on $\psi$)

- when $\psi = \theta$

$$
\pi(\theta \,|\, x, y) = \frac{\pi(\theta \,|\, x) m(y \,|\, \theta, x)}{m(y \,|\, x)} = \frac{\pi(\theta \,|\, x) g_\theta(y)}{m(y \,|\, x)}
$$

## MAP (maximum a posteriori) inferences

- the values $\psi$ are ordered: $\psi_2$ is preferred at least as much as $\psi_1$ whenever $\pi_\Psi(\psi_1 \mid x) \leq \pi_\Psi(\psi_2 \mid x)$

- motivation from the discrete case, $\psi_2$ is preferred at least as much as $\psi_1$ whenever the posterior prob. of $\psi_2$ is at least as big as the posterior prob. of $\psi_1$

- essentially evidence is being measured here by posterior probabilities

**E**: *posterior mode* $\psi(x) = \arg\sup \pi_\Psi(\psi \,|\, x)$ with error measured by the size of the $\gamma$*-highest posterior density (hpd) region*

$$C_{\Psi,\gamma}(x) = \{\psi : G_\Psi(\pi_\Psi(\psi \,|\, x) \,|\, x) \geq 1 - \gamma\}$$

where $G_\Psi(\cdot \,|\, x)$ is the posterior cdf of $\pi_\Psi(\psi \,|\, x)$ so $\Pi_\Psi(C_{\Psi,\gamma}(x) \,|\, x) \geq \gamma$

- how to choose $\gamma$? better than $\gamma$-likelihood regions because $\gamma$ is a probability here

**H**: to assess $H_0 = \{\psi_0\}$ compute (Bayesian p-value)

$$G_\Psi(\pi_\Psi(\psi_0 \,|\, x) \,|\, x) = \Pi_\Psi(\pi_\Psi(\psi \,|\, x) \leq \pi_\Psi(\psi_0 \,|\, x) \,|\, x)$$

and if this is small conclude evidence against (and no separate measure of the strength of the evidence)

- how small for evidence against?

**Example** *location normal*

- $\mu(x) = \mu_x = 7.23$

$$C_{\Psi,0.95}(x) = \mu(x) \pm 1.96\tau_x = [6.40, 8.06]$$

is the 0.95-hpd interval for $\mu$

- assess $H_0 = \{7\}$ then $G_\Psi(\pi_\Psi(7 \,|\, x) \,|\, x) > 0.05$ and so no evidence against

- in general there are two problems with MAP inferences with (2) more serious than (1)

**(1)** the inferences are not invariant under reparameterizations in the continuous case for if $\Xi : \Psi \overset{1-1, onto, smooth}{\longrightarrow} \Xi$ then posterior of $\xi = \Xi(\psi)$ is

$$\pi_\Xi(\xi \,|\, x) = \pi_\Psi(\Xi^{-1}(\xi) \,|\, x) J_\Xi(\Xi^{-1}(\xi))$$

and $J_\Xi(\Xi^{-1}(\xi))$ is not constant when $\Xi$ is nonlinear so $\xi(x) \neq \Xi(\psi(x))$ in general

**Example** *location normal*

- $\xi = \Xi(\mu) = \mu^3$ so $\mu = \xi^{1/3}$ and $J_\Xi(\Xi^{-1}(\xi)) = |\xi|^{-2/3}/3$ so the posterior of $\xi$ is

$$\pi_\Xi(\xi \,|\, x) = \frac{|\xi|^{-2/3}}{3\tau_x}\varphi\left(\frac{\xi^{1/3} - \mu_x}{\tau_x}\right)$$

which has an infinite singularity at $\xi = 0$ but in any case $\xi(x) \neq \mu^3(x)$

**(2)** probabilities do not measure evidence