# Hypothesis testing 3: Two-sample tests

Y. Polyanskiy, D. Shah, J. Tsitsiklis

6.S077

2018

Outline:

- Recap (p-value)
- Two-sample tests: paired an unpaired
- $t$-**test:** same variance and different variance
- Testing equality of distributions: KS-test, $G$-test, qqplot
- **Testing for independence**
- Confounding

# Review

## Definition

Statistical hypotheses:

- $H$ : data $X_1, \ldots, X_n$ distributed according to $P \in \mathcal{C}_0$
- $K$ : data $X_1, \ldots, X_n$ distributed according to $P \in \mathcal{C}_1$

where $\mathcal{C}_0, \mathcal{C}_1$ are COLLECTIONS OF DISTRIBUTIONS.

**Remarks:**

- Find statistic $T(X_1, \ldots, X_n)$ with $\approx$ same dist. under all $P \in \mathcal{C}_0$
- Test: If $T \geq t_\alpha$ then REJECT
- $t_\alpha$ is chosen depending on required size:

$$\max_{P \in \mathcal{C}_0} P[T \geq t_\alpha] \leq \alpha\,.$$

- Alternatively, report $p$-value: If $T(x_1, \ldots, x_n) = t_{obs}$

$$p = \max_{P \in \mathcal{C}_0} P[T \geq t_{obs}]$$

(aka "probability of same or more extreme data under null")

# Comparison (two-sample) testing

Examples:

- A/B testing in marketing
- one algorithm vs another
- one stock vs another
- new drug vs placebo

Two principal situations:

- Paired data:
  - each client tries both products
  - each input is evaluated using both algorithms
  - each day both stocks are evaluated
- Unpaired data:
  - each client tries only one product
  - each input is evaluated on only one product
  - each day can probe only one stock

# Two-sample testing (paired data)

- Paired data:
  - each client tries both products
  - each input is evaluated using both algorithms
  - each day both stocks are evaluated
- Statistical formalism: $(X_i, Y_i) \overset{iid}{\sim} P_{X,Y}$

$$H : \mathbb{E}[X] \leq \mathbb{E}[Y] \qquad K : \mathbb{E}[X] > \mathbb{E}[Y]$$

- For paired data can always reduce to one-sample case
- E.g. for real-valued measurements: $Z_i \triangleq X_i - Y_i$

$$H : \mathbb{E}[Z] \leq 0 \qquad K : \mathbb{E}[Z] > 0$$

- Hence use $z$-test or $t$-test (or Wald test)
- Already had example before

# Two-sample testing (unpaired data)

- Unpaired data:
  - each client tries only one product
  - each input is evaluated on only one product
  - each day can probe only one stock
- Statistical formalism: $X_i \overset{iid}{\sim} P_X, Y_i \overset{iid}{\sim} P_Y$

$$H : \mathbb{E}[X] \leq \mathbb{E}[Y] \qquad K : \mathbb{E}[X] > \mathbb{E}[Y]$$

- What to do?
- General idea:

$$\frac{\hat{\mu}_X - \hat{\mu}_Y}{\widehat{se}} \quad \lessgtr \quad t_\alpha \quad \boxed{\text{ACCEPT/REJECT}}$$

(aka two-sample $t$-test)

# Two-sample $t$-test: groups with same variance

- Hypothesis testing:
  - $X_i \overset{iid}{\sim} P_X$, $n$ samples, $\mu_X = \mathbb{E}[X]$
  - $Y_i \overset{iid}{\sim} P_Y$, $m$ samples, $\mu_Y = \mathbb{E}[Y]$
  - **Assumption:** $\mathrm{Var}[X] = \mathrm{Var}[Y] =$
  - ... same (but unknown!) variance of
  - One-sided: $H : \mu_X \leq \mu_Y$ vs $K :$
- General idea:

$$\frac{\hat{\mu}_X - \hat{\mu}_Y}{\widehat{se}} \lessgtr t_\alpha$$

ACC

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\bar{Y}_m = \frac{1}{m} \sum_{j=1}^m Y_j$$

- Let us try $T_0 = \bar{X}_n - \bar{Y}_m$ (sample means)

Need to normalize!

- Problem: What is $\mathrm{Var}[T_0]$?

$\mathrm{Var}[T_0] = \mathrm{Var}[\bar{X}_n] + \mathrm{Var}[\bar{Y}_m] = \sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)$

$\sigma^2$ unknown

- Unbiased estimator:

$\widehat{\sigma^2} = \frac{1}{n+m-2} \left( \sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{j=1}^m (Y_j - \bar{Y}_m)^2 \right)$

# Two-sample $t$-test: groups with same variance

**two-sample $t$-statistic (pooled variance)**

$$T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\widehat{\sigma^2}}}$$

$$\widehat{\sigma^2} = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \bar{Y}_m)^2\right)$$

and $\bar{X}_n$, $\bar{Y}_m$ are sample means.

- By LLN and CLT:

$$\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \to \mathcal{N}(0, \sigma^2)$$

$$\widehat{\sigma^2} \to \sigma^2$$

- ... Thus $T \to \mathcal{N}(0,1)$ as $n, m \to \infty$.

# Two-sample $t$-test: groups with same variance

- Hypothesis testing:
  - $X_i \overset{iid}{\sim} P_X$, $n$ samples, $\mu_X = \mathbb{E}[X]$
  - $Y_i \overset{iid}{\sim} P_Y$, $m$ samples, $\mu_Y = \mathbb{E}[Y]$
  - **Assumption:** $\mathrm{Var}[X] = \mathrm{Var}[Y] = \sigma^2$
  - One-sided: $H : \mu_X \leq \mu_Y \quad$ vs $\quad K : \mu_X > \mu_Y$
  - Two-sided: $H : \mu_X = \mu_Y \quad$ vs $\quad K : \mu_X \neq \mu_Y$
- Test statistic: $T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \frac{1}{\sqrt{\widehat{\sigma^2}}}$
- Thus select thresholds from approximating $T \approx \mathcal{N}(0,1)$:

| | | |
|---|---|---|
| $T > z_\alpha$ | REJECT | one-sided |
| $\lvert T \rvert > z_{\frac{\alpha}{2}}$ | REJECT | two-sided |

- If samples are $\approx \mathcal{N}$, then $\boxed{T \approx \texttt{scipy.stats.t.pdf}(\cdot, \texttt{n} + \texttt{m} - 2)}$

  (and thus replace $z_\alpha$ with $\alpha$-quantile of ...ditto...)
- Remember: pooled variance assumes **HOMOSCEDASTICITY**
- Example: signals (or patients) measured on the same noisy equipment.

# Two-sample $t$-test: general case

- Hypothesis testing:
  - $X_i \overset{iid}{\sim} P_X$, $n$ samples, $\mu_X = \mathbb{E}[X]$, $\mathrm{Var}[X] = \sigma_X^2$
  - $Y_i \overset{iid}{\sim} P_Y$, $m$ samples, $\mu_Y = \mathbb{E}[Y]$, $\mathrm{Var}[Y] = \sigma_Y^2$
  - One-sided: $H : \mu_X \leq \mu_Y$ vs $K : \mu_X > \mu_Y$
  - Two-sided: $H : \mu_X = \mu_Y$ vs $K : \mu_X \neq \mu_Y$
- Let us try $T_0 = \bar{X}_n - \bar{Y}_m$ (sample means) Need to normalize!
- Problem: What is $\mathrm{Var}[T_0]$?

  $\mathrm{Var}[T_0] = \mathrm{Var}[\bar{X}_n] + \mathrm{Var}[\bar{Y}_m] = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$ $\sigma$'s unknown
- Use unbiased estimators:

  $\widehat{\sigma_X^2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$

  $\widehat{\sigma_Y^2} = \frac{1}{m-1} \sum_{j=1}^{m} (Y_i - \bar{Y}_m)^2$

# Two-sample $t$-test: general case

- Hypothesis testing:
  - $X_i \stackrel{iid}{\sim} P_X$, $n$ samples, $\mu_X = \mathbb{E}[X]$, $\boxed{\mathrm{Var}[X] = \sigma_X^2}$
  - $Y_i \stackrel{iid}{\sim} P_Y$, $m$ samples, $\mu_Y = \mathbb{E}[Y]$, $\boxed{\mathrm{Var}[Y] = \sigma_Y^2}$
  - One-sided: $\boxed{H : \mu_X \leq \mu_Y \quad \text{vs} \quad K : \mu_X > \mu_Y}$
  - Two-sided: $\boxed{H : \mu_X = \mu_Y \quad \text{vs} \quad K : \mu_X \neq \mu_Y}$

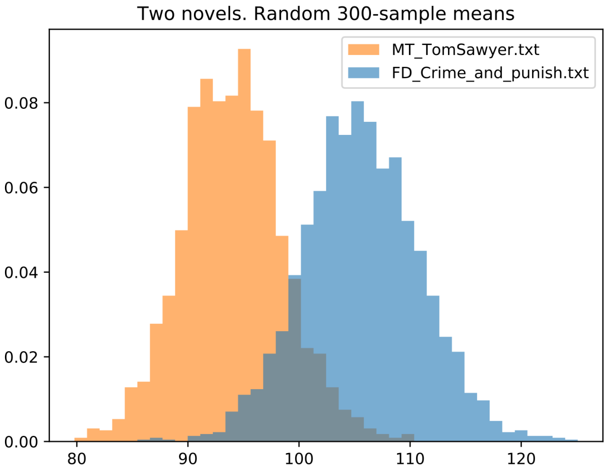
**two-sample $t$-statistic (unequal variance)**

$$T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\widehat{\sigma_X^2}}{n} + \frac{\widehat{\sigma_Y^2}}{m}}}$$

- Asymptotically normal: $T \approx \mathcal{N}(0,1)$, so use $z_\alpha$ or $z_{\alpha/2}$
- Small # samples: distribution unknown (even if $P_X, P_Y$ both $\mathcal{N}$).
- . . . aka Behrens-Fisher problem
- Welch correction: $T \approx$ Student-t with d.o.f.= (hard)
- Bootstrap: Simulate dist. of $T$ with $\tilde{X}_i \sim \mathcal{N}(0, \widehat{\sigma_X^2})$, $\tilde{Y}_i \sim \mathcal{N}(0, \widehat{\sigma_X^2})$

# Two-sample $t$-test: example

- Set
- Firs
- Ok, l




Sample-mean amplifies (and Gaussianizes) subtle differences.

# More than two groups, non-parametric tests

What we did not cover:

- Could have more than two groups
  - The null-hypothesis:
  $$H : \mu_1 = \mu_2 = \cdots = \mu_K$$
  - Test-statistic is (sort of):
  $$F = (\hat{\mu}_1 - \hat{\mu})^2 + \cdots + (\hat{\mu}_K - \hat{\mu})^2$$
  - Asymptotically $\chi^2()$-distributed under null.
  - Known as $F$-test
  - Such multiple-group problems have cool name: ANOVA
- Non-parametric tests:
  - $t$-tests are "parametric": exactly size-$\alpha$ only for Gaussian distributions.
  - Exactly size-$\alpha$ tests w/o assumptions?
  - Yes! And they are beautiful: Wilcoxon sum-rank tests
  - Key: sort $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$. If $P_X = P_Y$ then ranks of $X$'s and $Y$'s are uniformly distributed on $[n+m]$.

# Two-sample tests: beyond means

- Sometimes we may not be interested in means
  (e.g. data non-numerical)
- ... but still want to know if there is some effect
- Typical setting:
  - $X_i \overset{iid}{\sim} P_X$, $n$ samples
  - $Y_i \overset{iid}{\sim} P_Y$, $m$ samples
  - $\boxed{H : P_X = P_Y \quad \text{vs} \quad K : P_X \neq P_Y}$
- **Two cases:** continuous and discrete data

# Equality of distributions: Kolmogorov-Smirnov

- Last time: How to test $X \sim P_0$ with given (cts) $P_0$.
- Main observation: $\sqrt{n} \cdot \sup_t |\hat{F}_X(t) - F_0(t)|$ has known distribution (under null)
- Setting:
  - $X_i \overset{iid}{\sim} P_X$, $n$ samples
  - $Y_i \overset{iid}{\sim} P_Y$, $m$ samples
  - $H : P_X = P_Y$ vs $K : P_X \neq P_Y$

**two-sample Kolmogorov-Smirnov statistic**

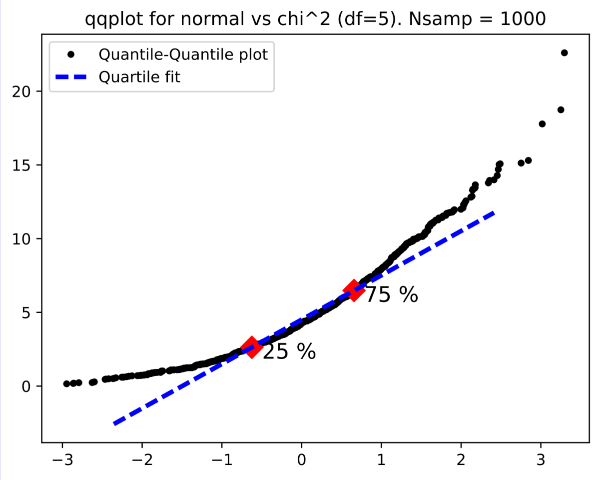
$$KS = \sqrt{\frac{nm}{n+m}} \cdot \sup_t |\hat{F}_X(t) - \hat{F}_Y(t)|$$

- Known distribution independent (!) of $P_X = P_Y$

  ...so just simulate on uniform to get $p$-value!
- Analytical formulae for $n, m \to \infty$. Use:

$$\texttt{scipy.stats.ks\_2samp}(\texttt{x\_samp}, \texttt{y\_samp})$$

# Equality of distributions: Quantile-Quantile plots



# Equality of discrete distributions

- Setting:
  - $X_i \overset{iid}{\sim} P_X$, $n$ samples
  - $Y_i \overset{iid}{\sim} P_Y$, $m$ samples
  - $H : P_X = P_Y$ vs $K : P_X \neq P_Y$
- **discrete $X$'s and $Y$'s**. Example:
  - Two hospitals
  - Hospital 1 sample: Cured, Cured, Died, . . . , Cured
  - Hospital 2 sample: Cured, Cured, Cured, . . . , Died
  - Summarize data in table:

| | Hospital 1 | Hospital 2 |
|---|---|---|
| Died | 3 | 10 |
| Cured | 33 | 54 |

  - Question: Columns generated by the same dist?
- Restate as follows:
  - New data: $(U_i, V_i)$ with $U \in \{\text{Died}, \text{Cured}\}$, $V \in \{1, 2\}$
  - Assume $(U_i, V_i) \overset{iid}{\sim} P_{U,V}$ and have $n + m$ such samples*
  - $H : U \perp\!\!\!\perp V$ vs $K : U \not\!\perp\!\!\!\perp V$
  - *subtlety: Orig. question was not symmetric in $U, V$
    (had samples $P_{U|V=1}$ and $P_{U|V=2}$) iid approx ok for $n, m \gg 1$

# G-test for independence

- New problem
  - $(U_i, V_i) \overset{iid}{\sim} P_{U,V}$, $\ell$-samples
  - $U$ is $t$-valued, i.e. $U \in [t] \triangleq \{1, \ldots, t\}$
  - $V$ is $s$-valued, i.e. $U \in [s] \triangleq \{1, \ldots, s\}$
  - $H : U \perp\!\!\!\perp V$ vs $K : U \not\!\perp\!\!\!\perp V$
- Recall generalized likelihood ratio test:

**The $G$-statistic (general)**

$$
\begin{aligned}
G &\triangleq -2 \log \frac{P_0^*(x_1, \ldots, x_n)}{P_1^*(x_1, \ldots, x_n)} \\
P_0^*(x_1, \ldots, x_n) &= \max_{P \in \mathcal{C}_0} P(x_1, \ldots, x_n) \\
P_1^*(x_1, \ldots, x_n) &= \max_{P \in \mathcal{C}_0 \cup \mathcal{C}_1} P(x_1, \ldots, x_n)
\end{aligned}
$$

**The $G$-statistic (test for independence)**

$$
G \triangleq 2\ell D(\hat{P}_{U,V} \| \hat{P}_U \times \hat{P}_V)
$$

# G-test for independence

- New problem
  - $(U_i, V_i) \overset{iid}{\sim} P_{U,V}$, $\ell$-samples
  - $U$ is $t$-valued, i.e. $U \in [t] \triangleq \{1, \ldots, t\}$
  - $V$ is $s$-valued, i.e. $U \in [s] \triangleq \{1, \ldots, s\}$
  - $H : U \perp\!\!\!\perp V$ vs $K : U \not\!\perp\!\!\!\perp V$
- Resulting test:
  - Compute $G_{norm} \triangleq 2\ell D(\hat{P}_{U,V} \| \hat{P}_U \times \hat{P}_V)$
  - Compare to $(1-\alpha)$-quantile of $\chi^2((t-1)(s-1))$
  - Alternatively,

$$p\text{-value} = \mathbb{P}[\chi^2((t-1)(s-1)) > G_{norm}] \,.$$

  - Or `scipy.stats.chi2.sf(`$G_{norm}$`,df=`$(t-1)(s-1)$`)`

| | Hospital 1 | Hospital 2 |
|---|---|---|
| Died | 3 | 10 |
| Cured | 33 | 54 |

$p$-value $= 0.28$

# Alternative test

- For 2x2 case don't need to be so fancy
- Test: $X \sim \mathrm{Bino}(n, p_1), Y \sim \mathrm{Bino}(m, p_2)$ and null $H : p_1 = p_2$.
- Do the two-sided $t$-test:

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}}$$

  with $\hat{p}_1 = X/n$ and $\hat{p}_2 = Y/m$;
- Same data as before (3:33, 10:54).
- $p$-value: $p = 0.2595$ (using $T \approx \mathcal{N}(0, 1)$)
- $p$-value: $p = 0.262$ (Welch corrected)
- $p$-value: $p = 0.260 \pm 0.001$ Bootstrap 1: equal-mean Binomial $X, Y$
- $p$-value: $p = 0.261 \pm 0.005$ Bootstrap 2: equal-mean Normal $X, Y$
- So why fancy $G$-test?
- Because it also works for $r$ hospitals and $s$ outcomes.

# Testing quality of classifiers

## Comparing quality of classifiers

Consider the following problem:

- Test set of size $n$ is given
- Two predictors (classifiers) are tested
- The base one has $1\%$ error.
- The new one has $e\%$ error
- Question: What $e$ is significant (to declare new one is better)?

- Can form a 2x2 contingency table and run a G-test
- Some sample numbers:
  - For $n = 10000$ (MNIST, CIFAR) we have

$$e < 0.65\% \quad \text{or} \quad e > 1.45\%$$

  - For $n = 1000$ we have

$$e < 0.13\% \quad \text{or} \quad e > 2.6\%$$

are significant (at $p = 0.05$)

# Beware of two-sample tests

- We learned how to test comparative hypotheses.
- BIG ISSUE: Confounding in observational studies
- Observational vs controlled study.
  - Observational study: groups self-selected
  - Randomized controlled study: groups assigned
- Cartoon example: "drinking beer makes you bald"

| | Bald | Not bald |
|---|---|---|
| Drinks beer | 49% | 2% |
| no beer | 1% | 48% |

Confounding factor: gender

correlation $\neq$ causation

# Beware of two-sample tests

**BIG ISSUE:** Confounding in observational studies.

Observational vs controlled study.

- Observational study: groups self-selected
- Randomized controlled study: groups assigned

- Real example:
  - Quinn et al [Nature'1999]: "Myopia and ambient lighting at night"
  - Eyeball development vs infant night sleep

| Sleep condition | Fraction developing myopia |
|---|---|
| Darkness | 10% |
| Night light | 34% |
| Room light | 55% |

  - Good sample size: $n = 479$
  - Sound statistics ($p$-value $< 0.00001$)
  - Physiologically plausible: *"The duration of the daily light period has been shown to affect eye growth in chicks"*
  - Gwiazda et al [Nature'2000]: could not reproduce
  - ... but: myopic parents are more likely to leave night light on

# Beware of two-sample tests

**What can we do about confounding?**

- Use common sense:
  - E.g. want to learn about effect of third kid on women labor market
  - Cannot do R.C.T.
  - Note: families with two kids of same sex are more likely to have third (by $6\%$)
  - ... use this for checking if two groups have similar unemployment

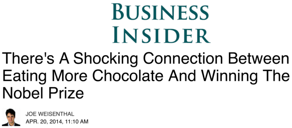
# Confounding

- Big area of research (Causal Inference)
- Rough idea: conditional independence testing
- If suspect relation between $X$ and $Y$ is confounded by $Z$ can test:

$$X \perp\!\!\!\perp Y | Z$$

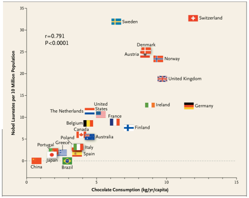
pro-term "controlling for $Z$"

**What can go wrong?** **Predicting $\ll$ Modeling**


BUSINESS INSIDER

There's A Shocking Connection Between Eating More Chocolate And Winning The Nobel Prize

JOE WEISENTHAL
APR. 20, 2014, 11:10 AM




The NEW ENGLAND JOURNAL of MEDICINE

$$\text{Choc} \to \text{Intell} \to \text{Nobel}$$

$$\begin{tikzcd} & \text{Choc} \\ \text{Wealth} \arrow[ur] \arrow[dr] & \\ & \text{Nobel} \end{tikzcd}$$

Need to infer a model/mechanism

**Can we? How?**