Chapter 3: Testing of Hypothesis

1 Basics of Hypothesis Testing

- Recall the basic concepts of Statistical Inference where we discussed: Statistical inference = Probability^{-1}.
- A curious question: Does my data come from a prescribed distribution, F?
- This is often called testing goodness of fit.
- Example: You roll a 6-sided die n times and observe $1, 3, 1, 6, 4, 2, 5, 3, \ldots$ Is this a fair die?

Motivating Example: Einstein's theory of Brownian motion[1]

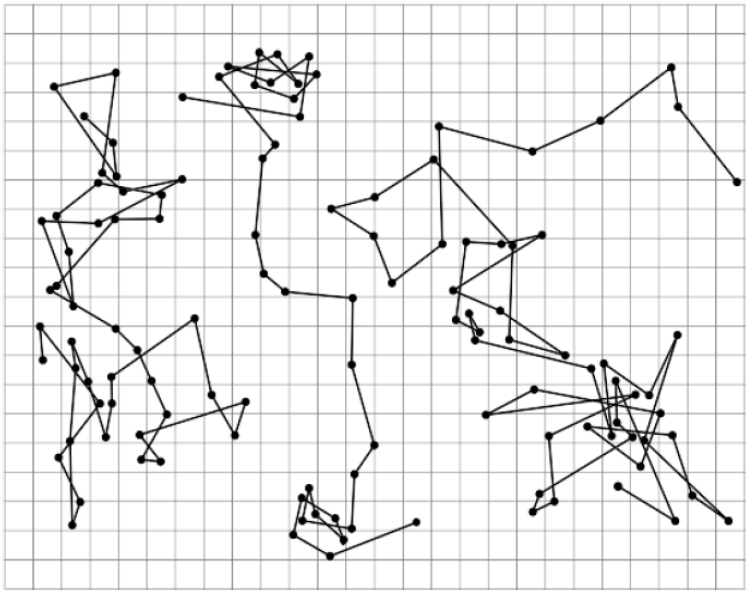


Figure 1: Motion of a tiny (radius $\approx 10^{-4}$ cm) particle suspended in water.

Albert Einstein (1905): $P_{t+\Delta t} \sim \mathcal{N}\left(P_t, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right)$, where $\sigma^2 = \frac{RT}{3\pi\eta r N_A}(\Delta t)$.

- P_t : position of particle at time t
- R: ideal gas constant

[1]Reading: A History of Random Processes (1968): https://www.jstor.org/stable/41133279

- T: absolute temperature
- η : viscosity of water
- r : radius of particle
- N_A : Avogadro's number

Jean Perrin (1909): Measured the position of a particle every 30 seconds to verify Einstein's theory (and to compute N_A). For his experiment, $\sigma^2 = 2.23 \times 10^{-7}$ cm^2.

- Does Perrin's data fit with Einstein's model?
- We will try to answer this question from testing of hypothesis viewpoint.

1.1 Null vs. Alternative Hypothesis

- A **hypothesis test** is a binary question about the data distribution. Our goal is to either accept a **null hypothesis** H_0 (which specifies something about this distribution) or to reject it in favor of an **alternative hypothesis** H_1.
- If H_0 (similarly H_1) completely specifies the probability distribution for the data, then the hypothesis is **simple**. Otherwise, it is **composite**.

1.2 Simple vs. Composite Hypothesis

Example 1.2.1. *Let $X_1, \ldots, X_6$ be the number of times we obtain 1 to 6 in n dice rolls. This null hypothesis is simple:*

$$H_0 : (X_1, \ldots, X_6) \sim \text{Multinomial}\left(n, \left(\frac{1}{6}, \ldots, \frac{1}{6}\right)\right).$$

We might wish to test this null hypothesis against the simple alternative hypothesis

$$H_1 : (X_1, \ldots, X_6) \sim \text{Multinomial}\left(n, \left(\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{2}{9}, \frac{2}{9}, \frac{2}{9}\right)\right),$$

or perhaps against the composite alternative hypothesis

$$H_1 : (X_1, \ldots, X_6) \sim \text{Multinomial}\left(n, (p_1, \ldots, p_6)\right) \quad \textit{for some } (p_1, \ldots, p_6) \neq \left(\frac{1}{6}, \ldots, \frac{1}{6}\right).$$

Example 1.2.2. *Let* $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots$ *be the displacement vectors* $P_{30} - P_0, P_{60} - P_{30}, P_{90} - P_{60}, \ldots$ *where* $P_t \in \mathbb{R}^2$ *is the position of a particle at time* t *in Perrin's experiment. Einstein's theory corresponds to the simple null hypothesis*

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{IID}{\sim} \mathcal{N}\left(0, 2.23 \times 10^{-7} I\right).$$

To test the theory qualitatively, but possibly allow for an error in Einstein's formula for σ^2*, we might test the composite null hypothesis*

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{IID}{\sim} \mathcal{N}\left(0, \sigma^2 I\right) \text{ for some } \sigma^2 > 0.$$

One can pose a number of different possible alternative hypotheses H_1 *to the above nulls (to be discussed later).*

1.3 Test statistics

A **test statistic** $T := T(X_1, \ldots, X_n)$ is any statistic such that extreme values (large or small) of T provide evidence against H_0.

Example 1.3.1. *Let* $X_1, \ldots, X_6$ *count the results from* n *dice rolls, and let*

$$T = \left(\frac{X_1}{n} - \frac{1}{6}\right)^2 + \ldots + \left(\frac{X_6}{n} - \frac{1}{6}\right)^2.$$

Large values of T *provide evidence against the null hypothesis of a fair die,*

$$H_0 : (X_1, \ldots, X_6) \sim \text{Multinomial}\left(n, \left(\frac{1}{6}, \ldots, \frac{1}{6}\right)\right).$$

Example 1.3.2. *Let* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *be the displacements from Perrin's experiment*[2]*. To test*

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{IID}{\sim} \mathcal{N}\left(0, 2.23 \times 10^{-7} I\right)$$

the following are possible test statistics:

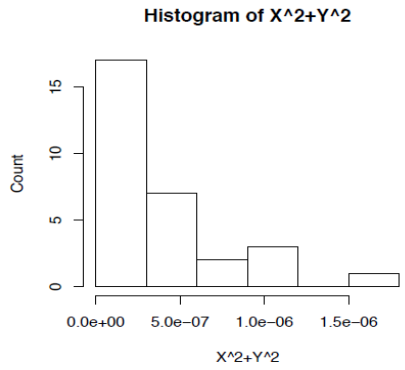
$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$
$$\bar{Y} = \frac{1}{n}(Y_1 + \ldots + Y_n)$$
$$V = \frac{1}{n}\left(X_1^2 + Y_1^2 + \ldots + X_n^2 + Y_n^2\right)$$

(Values of $\bar{X}$ *or* $\bar{Y}$ *much larger or smaller than 0 , or values of* V *much larger or smaller than* $2 \times 2.23 \times 10^{-7}$*, provide evidence against* H_0 *in favor of various alternatives* H_1*.)*

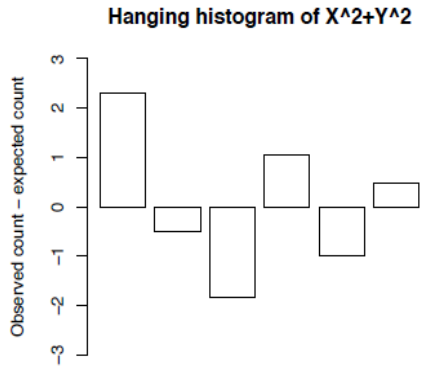
[2]Class Reading: https://aapt.scitation.org/doi/10.1119/1.2188962

1.4 Test statistics from histograms

Let $R_i = X_i^2 + Y_i^2$. Suppose we are interested in testing whether $R_1, \ldots, R_n$ are distributed as $2.23 \times 10^{-7}\chi_2^2$ (their distribution under H_0). We can plot a histogram of these values:



Deviations from $2.23 \times 10^{-7}\chi_2^2$ are better visualized by a hanging histogram, which plots $O_i - E_i$ where O_i is the observed count for bin i and E_i is the expected count under the $2.23 \times 10^{-7}\chi_2^2$ distribution:



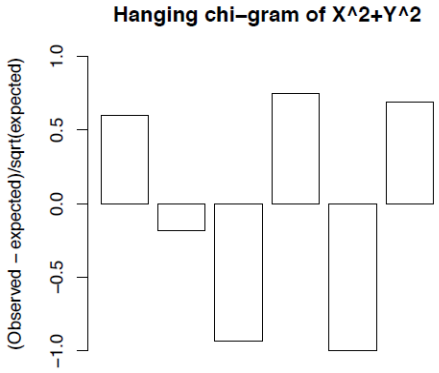
A test statistic can be $T = \sum_{i=1}^{6} (O_i - E_i)^2$.

Problem: Let p_i be the probability that the hypothesized chi-squared distribution assigns to bin i. If H_0 were true, then $O_i \sim \text{Binomial}(n, p_i)$ and $E_i = np_i = \mathbb{E}[O_i]$. So $\text{Var}[O_i] = \mathbb{E}\left[(O_i - E_i)^2\right] = np_i(1 - p_i)$. The variation in O_i is smaller, and scales approximately linearly with p_i, if p_i is close to 0. This might explain why the bars were smaller on the right side of the hanging histogram.

Solution: We can "stabilize the variance" by looking at

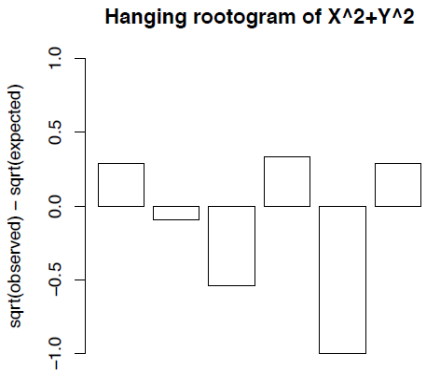
$$\frac{O_i - E_i}{\sqrt{E_i}} = \frac{O_i - E_i}{\sqrt{np_i}}$$

Or alternatively, we can look at $\sqrt{O_i} - \sqrt{E_i}$. (Taylor expansion of $\sqrt{x}$ around $x = E_i$ yields $\sqrt{O_i} - \sqrt{E_i} \approx \frac{1}{2\sqrt{E_i}}(O_i - E_i)$, so this has a similar effect as $\frac{O_i - E_i}{2\sqrt{E_i}}$ when $O_i - E_i$ is small.) The hanging chi-gram plots $\frac{O_i - E_i}{\sqrt{E_i}}$:



The test statistic $T = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$ is called **Pearson's chi-squared statistic for goodness of fit.**[3]

Tukey's hanging rootogram plots $\sqrt{O_i} - \sqrt{E_i}$:



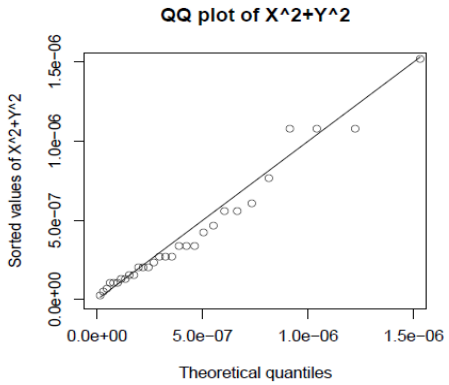
We may take as test statistic $T = \sum_{i=1}^{6} \left(\sqrt{O_i} - \sqrt{E_i}\right)^2$.

1.5 Test statistics from QQ plots

A **QQ plot** (or probability plot) compares the sorted values of $R_1, \ldots, R_n$ with the $\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}$ quantiles of the hypothesized $2.23 \times 10^{-7}\chi^2_2$ distribution:
Values close to the line $y = x$ indicate a good fit.

How do we get a test statistic from a QQ plot? One way is to take the maximum vertical deviation from the $y = x$ line: Let $R_{(1)} < \ldots < R_{(n)}$ be the sorted values of

[3] Xu, M., Zhang, D., & Wu, W. B. (2019). Pearson's chi-squared statistics: approximation theory and beyond. Biometrika, 106(3), 716-723.



$R_1, \ldots, R_n$. Take

$$T = \max_{i=1}^{n} \left| R_{(i)} - F^{-1}\left(\frac{i}{n+1}\right) \right|,$$

where F is the CDF of the $2.23 \times 10^{-7}\chi^2_2$ distribution so $F^{-1}(t)$ is its t^{th} quantile.

Problem: For values of R where the distribution has high density, the quantiles are closer together, so we expect a smaller vertical deviation. This explains why we see more vertical deviation in the upper right of the last QQ plot.

Solution: We may stabilize the spacings between quantiles by considering instead

$$T = \max_{i=1}^{n} \left| F\left(R_{(i)}\right) - \frac{i}{n+1} \right|.$$

This is almost the same as the **one-sample Kolmogorov-Smirnov (K-S) statistic**,

$$T_{KS} = \max_{i=1}^{n} \max\left(\left| F\left(R_{(i)}\right) - \frac{i}{n} \right|, \left| F\left(R_{(i)}\right) - \frac{i-1}{n} \right| \right).$$

(You can show $\frac{i-1}{n} < \frac{i}{n+1} < \frac{i}{n}$, and the difference between T and T_{KS} is negligible for large n.)

1.6 Null distributions and type I error

Supposing that we've picked our test statistic T, how large (or small) does T need to be, before we can safely assert that H_0 is false?

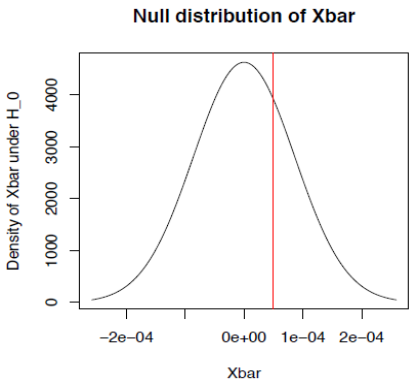
In most cases, we can never be 100% sure that H_0 is false. But we can compute T from the observed data and compare it with the sampling distribution of T if H_0 were true. This is called the **null distribution** of T.

Example 1.6.1. *Consider the following null hypothesis*

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{IID}{\sim} \mathcal{N}\left(0, 2.23 \times 10^{-7} I\right).$$

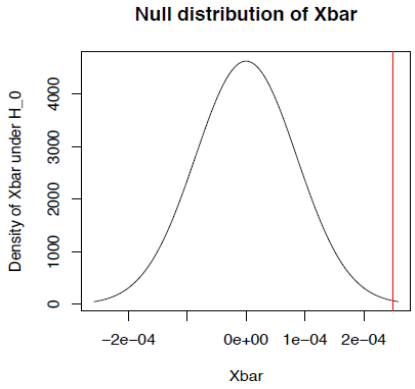
Under $H_0, \bar{X} \sim \mathcal{N}\left(0, 2.23 \times 10^{-7}/n\right)$. *This normal distribution is the null distribution of* $\bar{X}$.

Here's the PDF for the null distribution of $\bar{X}$, when $n = 30$:



If, for the observed data, $\bar{X} = 0.5 \times 10^{-4}$, this would not provide strong evidence against H_0. In this case, we might accept H_0.

Here's the PDF for the null distribution of $\bar{X}$, when $n = 30$:



If, for the observed data, $\bar{X} = 2.5 \times 10^{-4}$, this would provide strong evidence against H_0. In this case we might reject H_0.

The **rejection region** is the set of values of T for which we choose to reject H_0. The **acceptance region** is the set of values of T for which we choose to accept H_0.

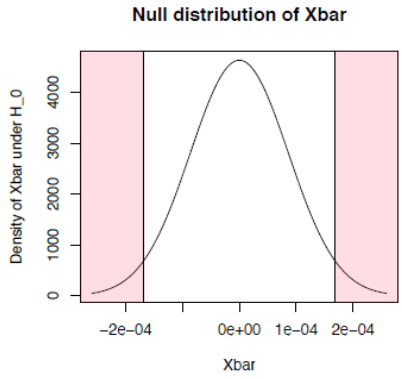
We choose the rejection region so as to control the probability of **type I error**:

$$\alpha = \mathbb{P}_{H_0}\left[\text{reject } H_0\right].$$

This value α is also called the **significance level** of the test.

If, under its null distribution, T belongs to the rejection region with probability α, then the test is level-α.

(Notation: For a simple null hypothesis H_0, we write $\mathbb{P}_{H_0}[\mathcal{E}]$ to denote the probability of event $\mathcal{E}$ under H_0, i.e., the probability of $\mathcal{E}$ if H_0 were true.)



Example 1.6.2. *A (two-sided) level-α test might reject H_0 when $\bar{X}$ falls in the above shaded regions. Mathematically, let z_α denote the $1-\alpha$ quantile, or "upper α point", of the distribution $\mathcal{N}(0,1)$. As $\bar{X} \sim \mathcal{N}\left(0, \sigma^2/n\right)$ under H_0 (where $\sigma^2 = 2.23 \times 10^{-7}$), the rejection region should be $\left(-\infty, -\frac{\sigma}{\sqrt{n}} \times z_{\alpha/2}\right] \cup \left[\frac{\sigma}{\sqrt{n}} \times z_{\alpha/2}, \infty\right)$.*

1.7 P values

The **p-value** is the smallest significance level at which your test would have rejected H_0.

For a one-sided test that rejects for large T, letting t_{obs} denote the value of T computed from the observed data, the p-value is $\mathbb{P}_{H_0}\left[T \geq t_{\text{obs}}\right]$.

For a two-sided test that rejects at the $\alpha/2$ and $1-\alpha/2$ quantiles of the null distribution of T, the p-value is 2 times the smaller of $\mathbb{P}_{H_0}\left[T \geq t_{\text{obs}}\right]$ and $\mathbb{P}_{H_0}\left[T \leq t_{\text{obs}}\right]$.

The p-value provides a quantitative measure of the extent to which the data supports (or does not support) H_0. It is preferable to report the exact p-value, rather than to just say "we rejected at level-0.05".

1.8 A word of caution

Accepting (or failing to reject) H_0 **does not** imply there is strong evidence that H_0 is true. Both of the following are possible:

- The particular test statistic you chose is not good at distinguishing the null hypothesis H_0 from the true distribution. Or equivalently, the true distribution is

not well-captured by the alternative H_1 that your test statistic is targeting. (For example, in **Perrin's data**, if there is significant drift in the y direction, you would not detect this using the test statistic $\bar{X}$.)

- You do not have enough data to reject H_0 at the significance level that you desire. In this case, your study might be **underpowered** (to be discussed later).

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β
Not rejected	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β

1.9 Determining the null distribution

To figure out the rejection region, we must understand the null distribution of the test statistic. There are three methods:

- Sometimes we can derive the null distribution exactly, for some of the previous problem where the test statistic is $\bar{X}$ and $X_1, \ldots, X_n$ are normally distributed under H_0.
- Sometimes we can derive an asymptotic approximation, using tools such as the CLT and continuous mapping theorem.
- When H_0 is simple, we can always obtain the null distribution by simulation.

1.10 Using an asymptotic null distribution

Example 1.10.1. *Let $(X_1, \ldots, X_6)$ denote the counts of 1 to 6 from n rolls of a die, and consider testing the simple null of a fair die*

$$H_0 : (X_1, \ldots, X_6) \sim \text{Multinomial}\left(n, \left(\frac{1}{6}, \ldots, \frac{1}{6}\right)\right)$$

using the test statistic

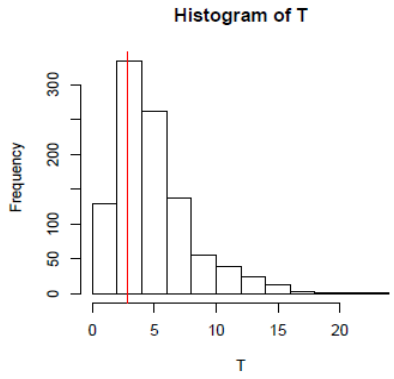
$$T = \left(\frac{X_1}{n} - \frac{1}{6}\right)^2 + \ldots + \left(\frac{X_6}{n} - \frac{1}{6}\right)^2.$$

It can be shown that for large n, T is approximately distributed as $\frac{1}{6n}\chi^2_5$.

*To perform an **asymptotic level-α test**, we may reject H_0 when t_{obs} exceeds $\frac{1}{6n}\chi^2_5(\alpha)$, where $\chi^2_n(\alpha)$ denotes the $1-\alpha$ quantile, or "upper α point", of the χ^2_n distribution.*

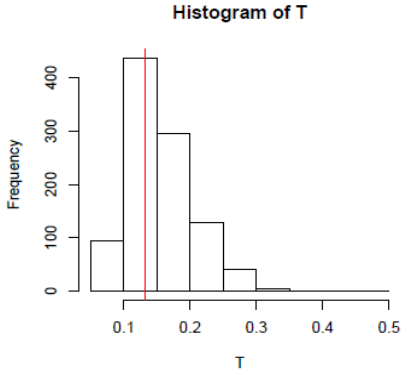
1.11 Using a simulated null distribution

Example 1.11.1. *Let T be Pearson's chi-squared statistic for goodness of fit for the values $X_1^2 + Y_1^2, \ldots, X_{30}^2 + Y_{30}^2$ from Perrin's experiments. We may simulate the null distribution of T :*



This shows the 1000 values of T across 1000 simulations. The observed value $t_{\text{obs}} = 2.83$ for Perrin's real data is in red.

Example 1.11.2. *Let T be the K-S statistic for $X_1^2 + Y_1^2, \ldots, X_{30}^2 + Y_{30}^2$. We may simulate the null distribution of T :*



The observed value $t_{\text{obs}} = 0.132$ for Perrin's real data is in red.

We obtain an approximate p-value as the fraction of simulated values of T larger than t_{obs}. (For a two-sided test, we would take either the fraction of simulated values of T larger than t_{obs} or smaller than t_{obs}, and multiply this by 2.)

For Perrin's data, the Pearson chi-squared p-value is 0.754, and the K-S p-value is 0.612. We accept H_0 in both cases, and neither test provides significant evidence against Einstein's theory of Brownian motion.

2 Simple alternatives and the Neyman-Pearson lemma

- Till now, we discussed a number of ways to construct test statistics for testing a simple null hypothesis, and we showed how to use the null distribution of the statistic to determine the rejection region so as to achieve the desired significance level.
- In this section, our goal is to answer the following question: Which test statistic should we use?
- The answer depends on the alternative hypothesis that we wish to distinguish from the null.

2.1 The Neyman-Pearson lemma

Let's focus on the problem of testing a simple null hypothesis H_0 against a simple alternative hypothesis H_1. We denote by

$$\beta = \mathbb{P}_{H_1}\left[\text{accept } H_0\right],$$

the probability of **type II error** – accepting the null H_0 when in fact the alternative H_1 is true. (Here, $\mathbb{P}_{H_1}[\mathcal{E}]$ denotes probability of an event $\mathcal{E}$ if H_1 is true.) Equivalently,

$$1 - \beta = \mathbb{P}_{H_1}\left[\text{reject } H_0\right],$$

is the probability of correctly rejecting H_0 when H_1 is true, which is called the **power** of the test against H_1.[4] When designing a hypothesis test for testing H_0 versus H_1, we have the following goal:

maximize: the power of the test against H_1
subject to: the significance level of the test under H_0 is at most α.

[4]Caution: Some books/papers use opposite notation and let β denote the power and $1 - \beta$ denote the probability of type II error. Make sure to double-check the meaning of the notation.

This is an example of a constrained optimization problem, which we can reason about in the following way: Suppose we observe data which are realizations of random variables $X_1, \ldots, X_n$. For notational convenience, let us denote by $\mathbf{X} = (X_1, \ldots, X_n)$ the entire data vector, and by $\mathbf{x} = (x_1, \ldots, x_n)$ a vector of possible values for $\mathbf{X}$. In the discrete case, suppose the hypotheses are

$$\begin{aligned} H_0 &: \mathbf{X} \text{ is distributed with joint PMF } f_0(\mathbf{x}) := f_0(x_1, \ldots, x_n), \\ H_1 &: \mathbf{X} \text{ is distributed with joint PMF } f_1(\mathbf{x}) := f_1(x_1, \ldots, x_n). \end{aligned}$$

Let $\mathcal{X}$ denote the set of all possible values of $\mathbf{X}$ under f_0 and f_1. To define the hypothesis test, for each $\mathbf{x} \in \mathcal{X}$, we must specify whether to accept or reject H_0 if the observed data is $\mathbf{x}$. In other words, we specify a rejection region $\mathcal{R} \subset \mathcal{X}$ such that we reject H_0 if the observed data belongs to $\mathcal{R}$ and we accept H_0 otherwise. Then the probability of rejecting H_0 if H_0 were true would be $\sum_{\mathbf{x}\in\mathcal{R}} f_0(\mathbf{x})$, the probability of rejecting H_0 if H_1 were true would be $\sum_{\mathbf{x}\in\mathcal{R}} f_1(\mathbf{x})$, and the above optimization problem is formalized as choosing the rejection region $\mathcal{R} \subset \mathcal{X}$ with the goal

$$\begin{aligned} &\text{maximize} \sum_{\mathbf{x}\in\mathcal{R}} f_1(\mathbf{x}) \\ &\text{subject to} \sum_{\mathbf{x}\in\mathcal{R}} f_0(\mathbf{x}) \leq \alpha. \end{aligned}$$

The continuous case is similar: suppose the hypotheses are

$$\begin{aligned} H_0 &: \mathbf{X} \text{ is distributed with joint PDF } f_0(\mathbf{x}) := f_0(x_1, \ldots, x_n), \\ H_1 &: \mathbf{X} \text{ is distributed with joint PDF } f_1(\mathbf{x}) := f_1(x_1, \ldots, x_n). \end{aligned}$$

We define a hypothesis test by defining the region $\mathcal{R} \subset \mathbb{R}^n$ such that we reject H_0 if and only if the observed data $\mathbf{x}$ belongs to $\mathcal{R}$. The above optimization problem is to choose $\mathcal{R} \subset \mathbb{R}^n$ with the goal

$$\begin{aligned} &\text{maximize} \int_{\mathcal{R}} f_1(\mathbf{x}) dx_1 \ldots dx_n \\ &\text{subject to} \int_{\mathcal{R}} f_0(\mathbf{x}) dx_1 \ldots dx_n \leq \alpha. \end{aligned}$$

In either the discrete or continuous case, what are the best points $\mathbf{x}$ to include in this rejection region $\mathcal{R}$? A moment's thought should convince you that $\mathcal{R}$ should consist of those points $\mathbf{x}$ corresponding to the smallest values of $\frac{f_0(\mathbf{x})}{f_1(\mathbf{x})}$, as these give the "smallest increase in type I error per unit increase of power". Another interpretation is that these are the points providing the strongest evidence in favor of H_1 over H_0. The statistic

$$L(\mathbf{X}) = \frac{f_0(\mathbf{X})}{f_1(\mathbf{X})}$$

is called the ***likelihood ratio statistic***, and the test that rejects for small values of $L(\mathbf{X})$ is called the ***likelihood ratio test***. The Neyman-Pearson lemma shows that the likelihood ratio test is the most powerful test of H_0 against H_1:

Theorem 2.1. *(Neyman-Pearson lemma). Let H_0 and H_1 be simple hypotheses (in which the data distributions are either both discrete or both continuous). For a constant $c > 0$, suppose that the likelihood ratio test which rejects H_0 when $L(\mathbf{x}) < c$ has significance level α. Then for any other test of H_0 with significance level at most α, its power against H_1 is at most the power of this likelihood ratio test.*

Proof. Consider the discrete case, and let $\mathcal{R} = \{\mathbf{x} : L(\mathbf{x}) < c\}$ be the rejection region of the likelihood ratio test. Note that among all subsets of $\mathcal{X}$, $\mathcal{R}$ maximizes the quantity

$$\sum_{x \in \mathcal{R}} \left(c f_1(\mathbf{x}) - f_0(\mathbf{x}) \right),$$

because $cf_1(\mathbf{x}) - f_0(\mathbf{x}) > 0$ for $\mathbf{x} \in \mathcal{R}$ and $cf_1(\mathbf{x}) - f_0(\mathbf{x}) \leq 0$ for $\mathbf{x} \notin \mathcal{R}$. Hence for any other test with significance level at most α, say with rejection region $\mathcal{R}'$,

$$\sum_{x \in \mathcal{R}} \left(c f_1(\mathbf{x}) - f_0(\mathbf{x}) \right) \geq \sum_{x \in \mathcal{R}'} \left(c f_1(\mathbf{x}) - f_0(\mathbf{x}) \right).$$

Rearranging the above, this implies

$$c \left(\sum_{x \in \mathcal{R}} f_1(\mathbf{x}) - \sum_{x \in \mathcal{R}'} f_1(\mathbf{x}) \right) \geq \sum_{x \in \mathcal{R}} f_0(\mathbf{x}) - \sum_{x \in \mathcal{R}'} f_0(\mathbf{x}) = \alpha - \sum_{x \in \mathcal{R}'} f_0(\mathbf{x}) \geq 0,$$

where the last inequality follows because $\sum_{x \in \mathcal{R}'} f_0(\mathbf{x})$ is the significance level of the test that rejects for $\mathbf{x} \in \mathcal{R}'$. Then $\sum_{x \in \mathcal{R}} f_1(\mathbf{x}) \geq \sum_{x \in \mathcal{R}'} f_1(\mathbf{x})$, i.e. the likelihood ratio test has power at least that of this other test. The proof in the continuous case is exactly the same, with all sums above replaced by integrals over $\mathcal{R}$ and $\mathcal{R}'$. □

2.2 Examples

Let's work out what the likelihood ratio test actually is for two simple examples.

Example 2.2.1. *Consider data $X_1, \ldots, X_n$ and the following null and alternative hypotheses:*

$$H_0 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0, 1)$$
$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, 1).$$

Here we assume μ is a known, specified value (not equal to 0), so that H_1 is a simple alternative hypothesis. The joint PDF *of $(X_1, \ldots, X_n)$ under H_0 is*

$$f_0(x_1,\ldots,x_n)=\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x_i^2}{2}}\right)=\left(\frac{1}{\sqrt{2\pi}}\right)^n\exp\left(-\frac{x_1^2+\ldots+x_n^2}{2}\right).$$

The joint PDF under H_1 is

$$f_1(x_1,\ldots,x_n)=\prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2}}\right)=\left(\frac{1}{\sqrt{2\pi}}\right)^n\exp\left(-\frac{(x_1-\mu)^2+\ldots+(x_n-\mu)^2}{2}\right).$$

Thus, the likelihood ratio statistic is

$$L(X_1,\ldots,X_n)=\frac{f_0(X_1,\ldots,X_n)}{f_1(X_1,\ldots,X_n)}=\exp\left(-\frac{X_1^2+\ldots+X_n^2}{2}+\frac{(X_1-\mu)^2+\ldots+(X_n-\mu)^2}{2}\right).$$

By expanding the squares and simplifying, we obtain

$$L(X_1,\ldots,X_n)=\exp\left(\frac{-2\mu(X_1+\ldots+X_n)+n\mu^2}{2}\right).$$

Suppose first that $\mu>0$. Then $L(X_1,\ldots,X_n)$ is a strictly decreasing function of the sample mean $\bar{X}=\frac{1}{n}(X_1+\ldots+X_n)$. Hence, rejecting for small values of $L(X_1,\ldots,X_n)$ is the same as rejecting for large values of $\bar{X}$. So the Neyman-Pearson lemma tells us that the most powerful test should reject when $\bar{X}>c$, for some threshold c. We pick c to ensure that the significance level is α under H_0: Since the null distribution of $\bar{X}$ is $\bar{X}\sim\mathcal{N}\left(0,\frac{1}{n}\right)$, c should be the $\frac{1}{\sqrt{n}}z_\alpha$ where z_α is the "upper α point" of the standard normal distribution.

Now suppose that $\mu<0$. Then $L(X_1,\ldots,X_n)$ is strictly increasing in $\bar{X}$, so rejecting for small $L(X_1,\ldots,X_n)$ is the same as rejecting for small $\bar{X}$. By the same argument as above, to ensure significance level α, the most powerful test rejects when $\bar{X}<-\frac{1}{\sqrt{n}}z_\alpha$.

Remark 2.1. *The most powerful test against the alternative $H_1: X_1,\ldots,X_n\sim\mathcal{N}(\mu,1)$ is the same for any $\mu>0$ (rejecting when $\bar{X}>\frac{1}{\sqrt{n}}z_\alpha$), and neither the test statistic nor the rejection region depends on the specific value of μ. This means that, in fact, this test is uniformly most powerful (UMP) against the (one-sided) composite alternative*

$$H_1: X_1,\ldots,X_n\overset{IID}{\sim}\mathcal{N}(\mu,1)\text{ for some }\mu>0.$$

On the other hand, the most powerful test is different for $\mu>0$ versus for $\mu<0$: one test rejects for large positive values of $\bar{X}$, and the other rejects for large negative values of $\bar{X}$. This implies that there does not exist a single most powerful test for the (two-sided) composite alternative

$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, 1) \text{ for some } \mu \neq 0.$$

Example 2.2.2. *Let $X_1, \ldots, X_n \in \{0, 1\}$ be the results of n flips of a coin, and consider the following null and alternative hypotheses:*

$$H_0 : X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$$
$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p).$$

Here we assume that $p \neq \frac{1}{2}$ is a known and specified value, so H_1 is simple. The joint PMF of $(X_1, \ldots, X_n)$ under H_0 and H_1 are, respectively,

$$f_0(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{2} = \frac{1}{2^n},$$
$$f_1(x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{x_1+\ldots+x_n}(1-p)^{n-x_1-\ldots-x_n} = (1-p)^n \left(\frac{p}{1-p}\right)^{x_1+\ldots+x_n}.$$

Thus, the likelihood ratio statistic is

$$L(X_1, \ldots, X_n) = \frac{f_0(X_1, \ldots, X_n)}{f_1(X_1, \ldots, X_n)} = \frac{1}{2^n(1-p)^n}\left(\frac{1-p}{p}\right)^{X_1+\ldots+X_n}.$$

First suppose $p > \frac{1}{2}$. Then $L(X_1, \ldots, X_n)$ is a decreasing function of $S = X_1 + \ldots + X_n$, so rejecting for small values of $L(X_1, \ldots, X_n)$ is the same as rejecting for large values of S. Hence, by the Neyman-Pearson lemma, the most powerful test rejects when $S > c$ for a constant c. We choose c to ensure significance level α : Under H_0, $S \sim$ Binomial $\left(n, \frac{1}{2}\right)$, so c should be the $1 - \alpha$ quantile of the Binomial $\left(n, \frac{1}{2}\right)$ *distribution. This test is the same for all $p > \frac{1}{2}$, so it is in fact uniformly most powerful against the composite alternative*

$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p) \text{ for some } p > \frac{1}{2}.$$

For $p < \frac{1}{2}$, $L(X_1, \ldots, X_n)$ is increasing in S, so the most powerful test rejects for $S < c$ and some constant c. To ensure significance level α, c should be the α quantile of the Binomial $\left(n, \frac{1}{2}\right)$ *distribution. This test is the same for all $p < \frac{1}{2}$, so it is uniformly most powerful against the compositive alternative*

$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p) \text{ for some } p < \frac{1}{2}.$$

Remark 2.2. *We have glossed over a detail, which is that when the distribution of the likelihood ratio statistic $L(\mathbf{X})$ is discrete under H_0, it might not be possible to choose c so that the significance level is exactly α. For instance, in the previous example, suppose we wish to achieve significance level $\alpha = 0.05$, and $n = 20$. For $S \sim \text{Binomial}\left(20, \frac{1}{2}\right)$, we have $\mathbb{P}[S \geq 15] = 0.021$ and $\mathbb{P}[S \geq 14] = 0.058$. So if we reject H_0 when $S \geq 14$, we do not achieve significance level $\leq \alpha$, and if we reject H_0 when $S \geq 15$, then we are too conservative.*

The theoretically correct solution is to perform a randomized test: Always reject H_0 when $S \geq 15$, always accept H_0 when $S \leq 13$, and reject H_0 with a certain probability when $S = 14$, where this probability is chosen to make the significance level exactly α. A more complete statement of the Neyman-Pearson lemma shows that this type of (possibly randomized) likelihood ratio test is most powerful among all randomized tests.

In practice, it might not be acceptable to use a randomized test. (We found the effects of this drug to be statistically significant because our statistical procedure told us to flip a coin, and our coin landed heads...) So we might take the more conservative option of just rejecting H_0 when $S \geq 15$.

3 Composite Hypotheses and the t-test

In this section we will discuss various hypothesis testing problems involving a composite null hypothesis and a composite alternative hypothesis.

3.1 Composite null and alternative hypotheses

To motivate the discussion, consider the following examples:

Example 3.1.1. *Suppose there are 80 students in a class taking MATH350 course. A diagnostic exam is administered at the start of the quarter, and a comparable exam is administered at the end of the quarter. Did the course MATH350 improve students' knowledge of statistics?*

Let X_i be the difference in test scores for student i. There are various ways we can formulate the above question as a hypothesis test: If we believe a normal model for the X_i's, $X_1, \ldots, X_{80} \overset{IID}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, then we might formulate our question as the testing problem:

$$H_0 : \mu = 0$$
$$H_1 : \mu > 0.$$

Note that both the null and alternative hypotheses above are composite, because they do not specify the variance σ^2 (which is unknown). If we are not willing to make a

normality assumption, we might assume instead that $X_1, \ldots, X_{80}$ are IID with PDF f, and test

$$H_0 : f \text{ is symmetric around } 0$$
$$H_1 : f \text{ is symmetric around } \mu \text{ for some } \mu > 0$$

or maybe even drop the symmetry assumptions and test

$$H_0 : f \text{ has median } 0$$
$$H_1 : f \text{ has median } \mu \text{ for some } \mu > 0.$$

Which formulation we choose and the resulting test statistic we use may depend on our prior knowledge of how test scores are typically distributed and on visual inspection of the data. (for departures from normality, symmetry, etc.)

Example 3.1.2. *A friend criticizes the setup of the previous example: It's hard to make two exams that are equally difficult. What if the second exam just happened to be a bit easier?*

To address this criticism, we add a control group: We give 100 other students (who are not taking statistics courses this quarter) the same two exams at the start and end of the quarter. Let Y_i be the difference in test scores for student i of this control group. Again, if we believe a normal model $X_1, \ldots, X_{80} \overset{IID}{\sim} \mathcal{N}\left(\mu_X, \sigma^2\right)$ and $Y_1, \ldots, Y_{100} \overset{IID}{\sim} \mathcal{N}\left(\mu_Y, \sigma^2\right)$ (with the X's also independent of the Y's), then we might formulate the test as

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X > \mu_Y.$$

If we are not willing to assume normality, we might suppose instead that $X_1, \ldots, X_{80}$ are IID with PDF f and $Y_1, \ldots, Y_{100}$ are IID with PDF g, and test

$$H_0 : f = g$$
$$H_1 : f \text{ stochastically dominates } g.$$

(This alternative H_1 means that if $X \sim f$ and $Y \sim g$, then $\mathbb{P}[X \geq x] \geq \mathbb{P}[Y \geq x]$ for all $x \in \mathbb{R}$.) Again, how we formulate the testing problem depends on the modeling assumptions we are willing to make.

When testing a composite null hypothesis H_0 against a compositive alternative H_1, there is a probability of type I error associated to each data distribution $P \in H_0$ (the probability of rejecting H_0 if the true distribution were P) and a probability of type II error associated to each data distribution $P \in H_1$ (the probability of accepting H_0 if the true distribution were P). A test has **significance level** α if the maximum

probability of type I error for any $P \in H_0$ is α.

This means that to design a level-α test of H_0, we need to control the probability of type I error for every $P \in H_0$, and hence reason about the sampling distribution of our test statistic T under every such data distribution P. In general this can be very difficult, and a common simplifying strategy will be to find a test statistic T that has *the same* sampling distribution under every $P \in H_0$.

3.2 One-sample t-test

Assume $X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ for unknown μ and σ^2, and consider testing

$$H_0 : \mu = 0$$
$$H_1 : \mu > 0.$$

If σ^2 were fixed and known, then the uniformly most-powerful level-α test would reject for large values of $\bar{X}$. Specifically, it would reject when $\frac{\sqrt{n}\bar{X}}{\sigma} > z_\alpha$ (because when $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}\left(0, \sigma^2\right), \bar{X} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$ so $\frac{\sqrt{n}\bar{X}}{\sigma} \sim \mathcal{N}(0,1)$).
When σ^2 is unknown, a natural idea is to estimate σ^2 by the sample variance

$$S^2 = \frac{1}{n-1}\left(\left(X_1 - \bar{X}\right)^2 + \ldots + \left(X_n - \bar{X}\right)^2\right),$$

and to consider the test statistic

$$T = \frac{\sqrt{n}\bar{X}}{S}.$$

To derive the distribution of T under H_0, we first prove the following result:

Theorem 3.1. *Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, and let $\bar{X}$ and S^2 be the sample mean and sample variance (where S^2 is defined as above). Then S^2 is independent of $\bar{X}$ and distributed as $S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$.*

Proof. Note that changing the mean μ does not affect the distribution of S^2 and shifts the distribution of $\bar{X}$ by a constant value, which does not affect independence of S^2 and $\bar{X}$. So we may assume without loss of generality $\mu = 0$.

We first show independence of S^2 and $\bar{X}$. The entries of $\left(\bar{X}, X_1 - \bar{X}, \ldots, X_n - \bar{X}\right)$ are linear combinations of $X_1, \ldots, X_n$, so $\left(\bar{X}, X_1 - \bar{X}, \ldots, X_n - \bar{X}\right)$ has a multivariate normal distribution by Example 1.9 of Chapter 1. Let's compute

$$\text{Cov}\left[\bar{X}, X_1 - \bar{X}\right] = \text{Cov}\left[\bar{X}, X_1\right] - \text{Cov}[\bar{X}, \bar{X}].$$

By bilinearity of covariance and the fact that $\text{Cov}\left[X_j, X_1\right] = 0$ for all $j \geq 2$,

$$\begin{aligned}\text{Cov}\left[\bar{X}, X_1\right] &= \text{Cov}\left[\frac{1}{n}\sum_{j=1}^{n} X_j, X_1\right] \\ &= \frac{1}{n}\sum_{j=1}^{n}\text{Cov}\left[X_j, X_1\right] = \frac{1}{n}\text{Cov}\left[X_1, X_1\right] = \frac{1}{n}\text{Var}\left[X_1\right] = \frac{\sigma^2}{n}.\end{aligned}$$

Since $\bar{X} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$, $\text{Cov}[\bar{X}, \bar{X}] = \text{Var}[\bar{X}] = \frac{\sigma^2}{n}$ also. Then

$$\text{Cov}\left[\bar{X}, X_1 - \bar{X}\right] = 0.$$

Similarly $\text{Cov}\left[\bar{X}, X_i - \bar{X}\right] = 0$ for every $i = 2, \ldots, n$. By Theorem 1.2 from Chapter 1, this means $\bar{X}$ is independent of $\left(X_1 - \bar{X}, \ldots, X_n - \bar{X}\right)$, and so $\bar{X}$ is independent of S^2.

To compute the distribution of S^2, we may write

$$\begin{aligned}(n-1)S^2 &= \left(X_1 - \bar{X}\right)^2 + \ldots + \left(X_n - \bar{X}\right)^2 \\ &= \left(X_1^2 - 2X_1\bar{X} + \bar{X}^2\right) + \ldots + \left(X_n^2 - 2X_n\bar{X} + \bar{X}^2\right) \\ &= X_1^2 + \ldots + X_n^2 - 2\left(X_1 + \ldots + X_n\right)\bar{X} + n\bar{X}^2 \\ &= \left(X_1^2 + \ldots + X_n^2\right) - 2n\bar{X}^2 + n\bar{X}^2 \\ &= \left(X_1^2 + \ldots + X_n^2\right) - n\bar{X}^2.\end{aligned}$$

Letting $U = (n-1)S^2/\sigma^2$, $W = \left(X_1^2 + \ldots + X_n^2\right)/\sigma^2$, and $V = n\bar{X}^2/\sigma^2$, this says $W = U + V$. We showed S^2 is independent of $\bar{X}$, hence U is independent of V. Thus the MGF of W is the product of the MGFs of U and V:

$$M_W(t) = M_U(t)M_V(t).$$

Finally, note that each $X_i/\sigma \sim \mathcal{N}(0,1)$, so $W \sim \chi^2_n$. Also, $\sqrt{n}\bar{X}/\sigma \sim \mathcal{N}(0,1)$, so $V = (\sqrt{n}\bar{X}/\sigma)^2 \sim \chi^2_1$. This means that the MGF of U is, for any $t < \frac{1}{2}$,

$$M_U(t) = \frac{M_W(t)}{M_V(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2},$$

which is the MGF of the χ^2_{n-1} distribution. So $U \sim \chi^2_{n-1}$, and $S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$. □

Remark 3.1. *Previously, we claimed if* $W_1, \ldots, W_6 \overset{IID}{\sim} \mathcal{N}(0,1)$, *then* $(W_1 - \bar{W})^2 + \ldots + \left(W_6 - \bar{W}\right)^2 \sim \chi^2_5$. *The above theorem verifies this. The theorem also explains why we often define* S^2 *with the normalization* $\frac{1}{n-1}$ *rather than* $\frac{1}{n}$*: As the expectation*

of a χ^2_{n-1} random variable is $n-1, \mathbb{E}\left[S^2\right] = \sigma^2$ so S^2 is an unbiased estimator for σ^2. Returning to our test statistic

$$T = \frac{\sqrt{n}\bar{X}}{S} = \frac{\sqrt{n}\bar{X}/\sigma}{S/\sigma},$$

we observe that by Theorem 3.1, for $\mu = 0$ and any value of $\sigma^2 > 0$,

$$\frac{\sqrt{n}\bar{X}}{\sigma} \sim \mathcal{N}(0,1), \quad \frac{S^2}{\sigma^2} \sim \frac{1}{n-1}\chi^2_{n-1},$$

and these are independent. Hence the distribution of T does not depend on σ, so it is the same under all $P \in H_0$. We give this distribution a name:

Definition 3.1. *If $Z \sim \mathcal{N}(0,1)$, $U \sim \chi^2_n$, and Z and U are independent, then the distribution of $\frac{Z}{\sqrt{\frac{1}{n}U}}$ is called the t-distribution with n degrees of freedom, denoted by t_n.*

So under H_0, $T \sim t_{n-1}$. Letting $t_{n-1}(\alpha)$ denote the upper α point (or $1-\alpha$ quantile) of the distribution t_{n-1}, the test that rejects for $T > t_{n-1}(\alpha)$ is called the **one-sample t-test**.

Remark 3.2. *The one-sample t-test is often used in paired two-sample settings, such as Example 3.1.1. There, we actually have two paired samples - the before and after test scores of each student - and we perform the test by first taking the differences of these paired values. In such settings, the test is often called the **paired two-sample t-test**, although the statistical procedure is really just a test for one set of IID observations.*

4 Two-sample t-test and signed rank test

4.1 Two-sample t-test

Consider the setting of two independent samples $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}\left(\mu_X, \sigma^2\right)$ and $Y_1, \ldots, Y_m \overset{\text{IID}}{\sim} \mathcal{N}\left(\mu_Y, \sigma^2\right)$, as in Example 3.1.2. Here μ_X, μ_Y, σ^2 are all unknown; note that we are assuming (for now) a common variance σ^2 for both samples. For the testing problem

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X > \mu_Y$$

a natural idea is to reject H_0 for large values of $\bar{X} - \bar{Y}$. Observe that $\bar{X} \sim \mathcal{N}\left(\mu_X, \frac{\sigma^2}{n}\right)$, $-\bar{Y} \sim \mathcal{N}\left(-\mu_Y, \frac{\sigma^2}{m}\right)$, and these are independent. Then their sum is distributed [5] as

[5] Recall, that if $X \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$ and $Y \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$ are independent, then $X + Y \sim \mathcal{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$.

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right).$$

Under $H_0,\ \mu_X - \mu_Y = 0$, so $\frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}} \sim \mathcal{N}(0,1)$. If σ^2 were known, then a level-α test based on $\bar{X} - \bar{Y}$ would reject when

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} > z(\alpha).$$

Since σ^2 is unknown, we estimate it from the data. We may use both the X_i's and Y_i's to estimate σ^2 by taking the **pooled sample variance**

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

and take as a test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

To derive the null distribution and rejection threshold for T, we may rewrite this as

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \Bigg/ \sqrt{S_p^2/\sigma^2}.$$

By Theorem 3.1 (and by independence of the two samples),
$\bar{X},\ \bar{Y},\ \sum_i (X_i - \bar{X})^2,\ \sum_j (Y_j - \bar{Y})^2$ are all independent, with the last two quantities distributed as $\sigma^2\chi^2_{n-1}$ and $\sigma^2\chi^2_{m-1}$. Then under H_0,

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim \mathcal{N}(0,1), \quad \frac{S_p^2}{\sigma^2} \sim \frac{1}{m+n-2}\chi^2_{m+n-2},$$

and these are independent. So the distribution of T is the same for all data distributions $P \in H_0$ and is given by

$$T \sim t_{m+n-2}.$$

The test that rejects H_0 when $T > t_{m+n-2}(\alpha)$ (the upper α point of the t_{m+n-2} distribution) is called the **two-sample t-test**.

Remark 4.1. *The assumption of common variance σ^2 for the two samples is oftentimes problematic (and violated) in practice. If we assume instead that $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}\left(\mu_X, \sigma_X^2\right)$ and $Y_1, \ldots, Y_m \overset{IID}{\sim} \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$ for possibly different values of σ_X^2 and σ_Y^2, then $\mathrm{Var}(\bar{X} - \bar{Y}) = \frac{1}{n}\sigma_X^2 + \frac{1}{m}\sigma_Y^2$, and we may estimate this by $\frac{1}{n}S_X^2 + \frac{1}{m}S_Y^2$, where S_X^2 and S_Y^2 are the sample variances of the two samples. Then we may use the test statistic*

$$T_{welch} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n}S_X^2 + \frac{1}{m}S_Y^2}}.$$

The distribution of T_{welch} under H_0 is no longer exactly a t distribution, but it was shown by Welch (1947)[6] to be close to the t distribution with

$$\frac{\left(S_X^2/n + S_Y^2/m\right)^2}{\left(S_X^2/n\right)^2/(n-1) + \left(S_Y^2/m\right)^2/(m-1)}$$

degrees of freedom. The test that rejects when T_{welch} exceeds the upper α point of this t distribution is called ***Welch's t-test*** *or the* ***unequal variances t-test****.*

4.2 Wilcoxon signed rank test (Non-parametric Test)

Let's return to the one-sample setting $X_1, \ldots, X_n \overset{IID}{\sim} f$, where we drop the normality assumption and only wish to test

$$H_0 : f \text{ is symmetric about } 0$$
$$H_1 : f \text{ is symmetric about } \mu \text{ for some } \mu > 0.$$

Because the shape of f is arbitrary under H_0, the distribution of the t-statistic is no longer the same under every data distribution $P \in H_0$ – in particular, it can be very far from t_{n-1} if n is moderately small and f is heavy-tailed. We consider instead the **signed rank statistic** W_+, defined in the following way:

1. Sort $|X_1|, |X_2|, \ldots, |X_n|$ in increasing order. Assign the smallest value (closest to zero) a rank of 1, the next smallest value a rank of 2, etc., and the largest value a rank of n.

2. Define W_+ as the sum of the ranks corresponding to only the positive values of $X_1, \ldots, X_n$.

As an example, suppose we have four observations $X_1 = 2, X_2 = -4, X_3 = -1, X_4 = 10$. Then the ranks of these four observations would be $2, 3, 1, 4$. Observations X_1 and X_4

[6] https://www.jstor.org/stable/2332510

are positive, so $W_+ = 2 + 4 = 6$.

We expect W_+ to be larger under H_1 than under H_0, because high-rank observations are more likely to be positive under H_1. The test that rejects for large W_+ is called **Wilcoxon's signed rank test**. The following theorem states that W_+ has the same distribution under every $P \in H_0$, and provides a method for determining the null distribution and rejection threshold for W_+ when n is large. (When n is small, we can determine the exact null distribution of W_+ by computing W_+ for all 2^n possible combinations of + and − signs for the ranked data.)

Theorem 4.1. *The distribution of W_+ is the same for every PDF f that is symmetric about 0. For large n, this distribution is approximately $\mathcal{N}\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$. (More formally, $\sqrt{\frac{24}{n(n+1)(2n+1)}}\left(W_+ - \frac{n(n+1)}{4}\right) \to \mathcal{N}(0,1)$ in distribution as $n \to \infty$.)*

Proof. We'll show that the distribution of W_+ is the same for every f, and that

$$\mathbb{E}[W_+] = \frac{n(n+1)}{4} \text{ and } \operatorname{Var}[W_+] = \frac{n(n+1)(2n+1)}{24}.$$

We'll provide only a heuristic explanation of why W_+ is asymptotically normal.

Let $f_0(x_1, \ldots, x_n) = \prod_{i=1}^n f(x_i)$ be the joint PDF of the data. By symmetry of f about 0, $f_0(\pm x_1, \ldots, \pm x_n)$ is the same for each of the 2^n combinations of +/− signs. This implies, conditional on $|X_1|, \ldots, |X_n|$, the signs of $X_1, \ldots, X_n$ are independent and each equal to + or − with probability $\frac{1}{2}$. Then, letting $I_k = 1$ if the value with rank k is positive and $I_k = 0$ if it is negative, $I_1, \ldots, I_n \overset{IID}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$ for any PDF f that is symmetric about 0.

The signed rank statistic is

$$W_+ = \sum_{k=1}^n k I_k.$$

Since $I_1, \ldots, I_n$ have the same distribution under any symmetric PDF f about 0, the distribution of W_+ is the same for all such PDFs f. We compute

$$\mathbb{E}[W_+] = \sum_{k=1}^n k\mathbb{E}[I_k] = \frac{1}{2}\sum_{k=1}^n k = \frac{n(n+1)}{4},$$
$$\operatorname{Var}[W_+] = \sum_{k=1}^n \operatorname{Var}[kI_k] = \sum_{k=1}^n k^2 \operatorname{Var}[I_k] = \frac{1}{4}\sum_{k=1}^n k^2 = \frac{n(n+1)(2n+1)}{24},$$

where the computation for variance uses that $I_1, \ldots, I_n$ are independent.

To explain why W_+ is approximately normally distributed, define the **empirical CDF** of $|X_1|, \ldots, |X_n|$ by

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{|X_i| \leq t\}.$$

($F_n(t)$ is the fraction of values of $|X_i|$ that are at most t.) Then the rank associated with X_i is exactly $nF_n(|X_i|)$, so

$$W_+ = \sum_{i=1}^{n} nF_n(|X_i|)\, \mathbb{1}\{X_i > 0\}.$$

When n is large, one may show that $F_n(t)$ is, with high probabiliy, close to the true CDF $F(t)$ of $|X_i|$ for every $t \in \mathbb{R}$, and hence that the difference between $W_+$and

$$\tilde{W}_+ = \sum_{i=1}^{n} nF(|X_i|)\, \mathbb{1}\{X_i > 0\}$$

is negligible. But $\tilde{W}$ is just the sum of IID random variables $Y_i := nF(|X_i|)\, \mathbb{1}\{X_i > 0\}$, and hence asymptotically normally distributed by the CLT. □

5 Rank sum test and Permutation tests

5.1 Rank sum test

The idea of converting observed data values to just their ranks, so as to deal with heavy-tailed data and deviations from normality, can be extended to the two-sample setting. Consider two independent samples $X_1, \ldots, X_n \overset{IID}{\sim} f$ and $Y_1, \ldots, Y_m \overset{IID}{\sim} g$, where f and g are two arbitrary PDFs, and the testing problem

$$H_0 : f = g$$
$$H_1 : f \text{ stochastically dominates } g.$$

(Recall from Section 3 that this alternative is one way of saying that values drawn from f "tend to be larger" than values drawn from g.)

The **rank-sum statistic** T_Y is defined as follows:

1. Consider the **pooled sample** of all observations $X_1, \ldots, X_n, Y_1, \ldots, Y_m$. Sort these $m+n$ values in increasing order. Assign the smallest a rank of 1, the next smallest a rank of 2, etc., and the largest a rank of $m+n$.

2. Define T_Y as the sum of the ranks corresponding to only the Y_i values, i.e., the values from only the second sample[7].

We expect T_Y to be smaller under H_1 than under H_0, because under H_1 the values of Y_i tend to have smaller ranks. The test that rejects for small values of T_Y is called the **Wilcoxon rank-sum test**, known alternatively as the Mann-Whitney U-test or the Mann-Whitney Wilcoxon test.

(If we are testing a general two-sided alternative $H_1' : f \neq g$ then we would reject for both large and small values of T_Y.)

The following theorem states that T_Y has the same distribution under every $P \in H_0$, and provides a method for determining the null distribution and rejection threshold when n and m are both large. (For small n and m, we can determine the exact null distribution of T_Y by computing T_Y for all $\begin{pmatrix} n+m \\ m \end{pmatrix}$ possible sets of ranks for the Y_i 's.)

Theorem 5.1. *The distribution of T_Y is the same under any PDF $f = g$. For large n and m, this distribution is approximately $\mathcal{N}\left(\frac{m(m+n+1)}{2}, \frac{mn(m+n+1)}{12}\right)$.*

We won't prove this result; let's just make the following comments:

- If $f = g$, then each ordering of $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ is equally likely. Since T_Y depends only on this ordering, its distribution must be the same under every PDF $f = g$
- Let $I_k = 1$ if the k^{th} largest value in $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ belongs to the second sample, and $I_k = 0$ otherwise. Then

$$T_Y = \sum_{k=1}^{m+n} kI_k.$$

Under H_0, I_k indicates whether the k^{th} "individual" is selected in a simple random sample of size m (without replacement) from a population of size $m + n$. Then the same computations as in Lecture 1 yield formulas for $\mathbb{E}[I_k]$, $\mathrm{Var}[I_k]$, and $\mathrm{Cov}[I_j, I_k]$. Applying linearity of expectation and bilinearity of covariance, we may obtain

$$\mathbb{E}[T_Y] = \frac{m(m+n+1)}{2} \text{ and } \mathrm{Var}[T_Y] = \frac{mn(m+n+1)}{12}$$

as in the above theorem. (Details are provided in Rice [8], Section 11.2.3 Theorem A and Section 7.3.1 Theorems A and B.)

[7] One may consider equivalently T_X (the sum of ranks of the X_i's) as $T_X + T_Y$ is a fixed constant.

[8] Rice, John A. Mathematical statistics and data analysis. Cengage Learning, 2006.

5.2 Permutation and randomization tests

The main idea behind the (one-sample) signed-rank test and the (two-sample) rank-sum test is to exploit a symmetry under H_0. For the signed-rank test, the symmetry is that it is equally likely to observe $\pm X_1, \ldots, \pm X_n$ for each of the 2^n combinations of $+/-$ signs. For the rank-sum test, the symmetry is that it is equally likely to observe each of the $(m+n)!$ permutations of the pooled sample $X_1, \ldots, X_n, Y_1, \ldots, Y_m$.

In fact, this idea of exploiting symmetry provides an alternative (and useful) simulation based method of obtaining a null distribution for any test statistic T for these problems:

Example 5.2.1. *Consider two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, and any test statistic $T(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$. (For concreteness, you can think about $T = \bar{X} - \bar{Y}$.) For a null hypothesis H_0 which specifies that all data from both samples are IID from a common distribution, for example*

$$H_0 : X_1, \ldots, X_n, Y_1, \ldots, Y_m \overset{IID}{\sim} f$$

*for an unknown PDF f, the **permutation null distribution** of T is the distribution of $T(X_1^*, \ldots, X_n^*, Y_1^*, \ldots, Y_m^*)$ when we fix the observed values $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ and let $(X_1^*, \ldots, X_n^*, Y_1^*, \ldots, Y_m^*)$ be a permutation of $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ chosen uniformly at random from the set of all $(m+n)!$ possible permutations. (For $T = \bar{X} - \bar{Y}$, what this effectively means is that we randomly choose n of the observations to be $X_1^*, \ldots, X_n^*$, set the remaining m observations to be $Y_1^*, \ldots, Y_m^*$, and compute $\bar{X}^* - \bar{Y}^*$.)*

Under H_0, each of these $(m+n)!$ possible values of T is equally likely to be observed. To perform a test that rejects for large values of T, we may use the following procedure:

1. *Randomly permute the pooled data B times (say $B = 10000$), and compute the value of T each time.*
2. *Compute an approximate p-value as the fraction of the B simulations where we obtained a value of T larger than t_{obs}, the value for the original (unpermuted) data. (Reject at level- α if this p-value is at most α.)*

For a two-sided test that rejects for both large and small values of T, we can compute the p-value by taking the fraction of simulations where T is larger than t_{obs} or the fraction where T is smaller than t_{obs} (whichever is smaller) and multiply by 2.

This is called a **permutation test** based on T. It is an example of a **conditional test**, because we are looking at the conditional distribution of the data under H_0 given

the set (but not the ordering) of their values.

The utility of this idea is that it may be applied to test statistics T where we do not understand its (unconditional) distribution under H_0, and where this distribution may vary for different PDFs $f = g$. Consider the following example:

Example 5.2.2. *Let $X_1, \ldots, X_n \in \mathcal{X}$ and $Y_1, \ldots, Y_m \in \mathcal{X}$ be two random samples of "objects" (e.g. images, websites, documents) represented in some data space $\mathcal{X}$. Suppose we have a function $d(x, y)$ that measures a "distance" between any two objects $x, y \in \mathcal{X}$.*

To test whether $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ appear to come from the same distribution, the following might be a reasonable test statistic:

$$T_1 = \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} d\left(X_i, Y_j\right) - \frac{1}{\binom{n}{2}} \sum_{1 \le i < i' \le n} d\left(X_i, X_{i'}\right) - \frac{1}{\binom{m}{2}} \sum_{1 \le j < j' \le m} d\left(Y_j, Y_{j'}\right).$$

In words, T_1 is twice the average distance between an object in sample 1 and an object in sample 2, minus the average distance between two objects in sample 1 and minus the average distance between two objects in sample 2 . So T_1 measures whether, on average, objects from the same sample are more similar to each other than objects from different samples.

Or we might consider a "nearest-neighbors" statistic: For each of the $m + n$ data values, look at the k other data values closest to it (as measured by the distance d) and count how many of these come from the same sample as itself. Let T_2 be the average of this count across all $m + n$ data points. So T_2 measures whether the k closest other objects tend to come from the same sample.

The distributions of T_1 and T_2 under H_0 may be difficult to understand theoretically and may depend on the unknown common distribution of $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, but we can still carry out a permutation test based on T_1 or on T_2.

A similar idea may be applied in the one-sample setting for testing the null hypothesis

$$H_0 : X_1, \ldots, X_n \overset{\text{IID}}{\sim} f, \text{ for some PDF } f \text{ symmetric about } 0$$

based on the symmetry underlying the Wilcoxon signed-rank test.

6 Hypothesis Testing for Categorical Data

Here, we introduce the generalized likelihood ratio test (GLRT) and explore applications to the analysis of categorical data.

6.1 GLRT for a simple null hypothesis

Let $\{f(x \mid \theta) : \theta \in \Omega\}$ be a parameteric model, and let $\theta_0 \in \Omega$ be a particular parameter value. For testing

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

the **generalized likelihood ratio test (GLRT)** rejects for small values of the test statistic

$$\Lambda = \frac{\text{lik}\,(\theta_0)}{\max\limits_{\theta \in \Omega} \text{lik}(\theta)},$$

where $\text{lik}(\theta)$ is the likelihood function. (In the case of IID samples $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$, $\text{lik}(\theta) = \prod_{i=1}^n f\,(X_i \mid \theta)$.) The numerator is the value of the likelihood at θ_0, and the denominator is the value of the likelihood at the MLE $\hat{\theta}$. The level- α GLRT rejects H_0 when $\Lambda \leq c$, where (as usual) c is chosen so that $\mathbb{P}_{H_0}[\Lambda \leq c]$ equals (or approximately equals) α.

Note that the GLRT differs from the likelihood ratio test discussed previously in the context of the Neyman-Pearson lemma, where the denominator was instead given by $\text{lik}\,(\theta_1)$ for a simple alternative $\theta = \theta_1$. The alternative H_1 above is not simple, and the GLRT replaces the denominator with the maximum value of the likelihood over all values of θ.

Example 6.1.1. *Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\theta, 1)$ and consider the problem of testing*

$$H_0 : \theta = 0$$
$$H_1 : \theta \neq 0.$$

The MLE for θ is $\hat{\theta} = \bar{X}$. We compute

$$\text{lik}(0) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{X_i^2}{2}}$$
$$\max_{\theta \in \mathbb{R}} \text{lik}(\theta) = \text{lik}(\hat{\theta}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i - \bar{X})^2}{2}}.$$

Then

$$\Lambda = \frac{\text{lik}(0)}{\max_{\theta\in\mathbb{R}} \text{lik}(\theta)} = \exp\left(-\sum_{i=1}^{n}\frac{X_i^2}{2} + \sum_{i=1}^{n}\frac{\left(X_i-\bar{X}\right)^2}{2}\right)$$
$$= \exp\left(-\sum_{i=1}^{n}\frac{X_i^2}{2} + \sum_{i=1}^{n}\frac{X_i^2 - 2X_i\bar{X}+\bar{X}^2}{2}\right) = \exp\left(-\frac{n}{2}\bar{X}^2\right).$$

Rejecting for small values of Λ is the same as rejecting for large values of $-2\log\Lambda = n\bar{X}^2$. Under H_0, $\sqrt{n}\bar{X} \sim \mathcal{N}(0,1)$, so $n\bar{X}^2 \sim \chi_1^2$. Then the GLRT rejects H_0 when $n\bar{X}^2 > \chi_1^2(\alpha)$, the upper-$\alpha$ point of the χ_1^2 distribution. (This is the same as rejecting when $|\bar{X}| > z(\alpha/2)/\sqrt{n}$, so the GLRT is equivalent to usual two-sided z-test based on $\bar{X}$.)

In general, the exact sampling distribution of $-2\log\Lambda$ under H_0 may not have a simple form as in the above example, but it may be approximated by a chi-squared distribution for large n:

Theorem 6.1. *Let $\{f(x \mid \theta) : \theta \in \Omega\}$ be a parametric model and let $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta_0)$. Suppose θ_0 is an interior point of Ω, and the regularity conditions of Theorems 1.1 and 2.1 of Chapter 2 (for consistency and asymptotic normality of the MLE) hold. Then*

$$-2\log\Lambda \to \chi_k^2$$

in distribution as $n \to \infty$, where $k = \dim\Omega$ is the dimension of Ω.

Hint. For simplicity, we consider only the case $k = 1$, so θ is a single parameter. Letting $l(\theta)$ denote the log-likelihood function and $\hat{\theta}$ denote the MLE,

$$-2\log\Lambda = -2l(\theta_0) + 2l(\hat{\theta}).$$

Applying a Taylor expansion of $l(\theta_0)$ around $\theta_0 = \hat{\theta}$,

$$l(\theta_0) \approx l(\hat{\theta}) + \left(\theta_0 - \hat{\theta}\right) l'(\hat{\theta}) + \frac{1}{2}\left(\theta_0 - \hat{\theta}\right)^2 l''(\hat{\theta}) \approx l(\hat{\theta}) - \frac{1}{2}nI(\theta_0)\left(\theta_0 - \hat{\theta}\right)^2,$$

where the second approximation uses $l'(\hat{\theta}) = 0$ and $l''(\hat{\theta}) \approx -nI(\hat{\theta}) \approx -nI(\theta_0)$. Then

$$-2\log\Lambda \approx nI(\theta_0)\left(\theta_0 - \hat{\theta}\right)^2.$$

$\sqrt{nI(\theta_0)}\left(\hat{\theta}-\theta_0\right) \to \mathcal{N}(0,1)$ in distribution by asymptotic normality of the MLE, so the continuous mapping theorem implies $-2\log\Lambda \approx nI(\theta_0)\left(\theta_0-\hat{\theta}\right)^2 \to \chi^2_1$ as desired. □

This theorem implies that an approximate level-α test is given by rejecting H_0 when $-2\log\Lambda > \chi^2_k(\alpha)$, the upper- α point of the χ^2_k distribution. The "dimension" k of Ω is the number of free parameters in the model or the number of parameters minus the number of independent constraints. For instance, in Example 6.1.1, there is a single parameter θ, so the dimension is 1. For a multinomial model with parameters $(p_1,\ldots,p_k)$, there are k parameters but they are constrained to sum to 1 , so the dimension is $k-1$.

6.2 GLRT for testing a sub-model

More generally, let $\Omega_0 \subset \Omega$ be a subset of the parameter space Ω, corresponding to a lower-dimensional sub-model. For testing

$$H_0 : \theta \in \Omega_0$$
$$H_1 : \theta \notin \Omega_0$$

the generalized likelihood ratio statistic is defined as

$$\Lambda = \frac{\max_{\theta\in\Omega_0}\mathrm{lik}(\theta)}{\max_{\theta\in\Omega}\mathrm{lik}(\theta)}.$$

In other words, Λ is the ratio of the values of the likelihood function evaluated at the MLE in the sub-model and at the MLE in the full-model.

For large n, under any $\theta_0 \in \Omega_0$, $-2\log\Lambda$ is approximately distributed as χ^2_k where k is the difference in dimensionality between Ω_0 and Ω, and an approximate level-α test rejects H_0 when $-2\log\Lambda > \chi^2_k(\alpha)$:

Theorem 6.2. *Let $\{f(x \mid \theta) : \theta \in \Omega\}$ be a parametric model, and let $X_1,\ldots,X_n \overset{IID}{\sim} f(x \mid \theta_0)$ where $\theta_0 \in \Omega_0$. Suppose θ_0 is an interior point of both Ω_0 and Ω, and the regularity conditions of Theorems 1.1 and 2.1 of Chapter 3 hold for both the full model $\{f(x \mid \theta) : \theta \in \Omega\}$ and the sub-model $\{f(x \mid \theta) : \theta \in \Omega_0\}$. Then*

$$-2\log\Lambda \to \chi^2_k$$

in distribution as $n \to \infty$, where $k = \dim\Omega - \dim\Omega_0$.

Figure 2: The Hardy-Weinberg theorem characterizes the distributions of genotype frequencies in populations that are not evolving, and is thus the fundamental null model for population genetics.

Example 6.2.1. *(Hardy-Weinberg equilibrium[9]). At a single diallelic locus in the genome with two possible alleles* A *and a, any individual can have genotype AA, Aa, or aa. If we randomly select n individuals from a population, we may model the numbers of individuals with these genotypes as* $(N_{AA}, N_{Aa}, N_{aa}) \sim \text{Multinomial}\,(n, (p_{AA}, p_{Aa}, p_{aa}))$.

When the alleles A *and a are present in the population with proportions* θ *and* $1-\theta$*, then under an assumption of random mating, quantitative genetics theory predicts that* p_{AA}*,* p_{Aa}*, and* p_{aa} *should be given by* $p_{AA} = \theta^2, p_{Aa} = 2\theta(1-\theta)$*, and* $p_{aa} = (1-\theta)^2$*-this is called the Hardy-Weinberg equilibrium. In practice we do not know* θ*, but we may still test the null hypothesis that Hardy-Weinberg equilibrium holds for some* θ*:*

$$H_0 : p_{AA} = \theta^2,\ p_{Aa} = 2\theta(1-\theta),\ p_{aa} = (1-\theta)^2 \text{ for some } \theta \in (0,1).$$

This null hypothesis corresponds to a 1-dimensional sub-model (with a single free parameter θ *) inside the 2-dimensional multinomial model (specified by general parameters* p_{AA}, p_{Aa}, p_{aa} *summing to 1). We may test* H_0 *using the GLRT:*
The multinomial likelihood is given by

$$l\,(p_{AA}, p_{Aa}, p_{aa}) = \binom{n}{N_{AA}, N_{Aa}, N_{aa}} p_{AA}^{N_{AA}} p_{Aa}^{N_{Aa}} p_{aa}^{N_{aa}}.$$

Letting $\hat{p}_{AA}, \hat{p}_{Aa}, \hat{p}_{aa}$ *denote the full-model MLEs and* $\hat{p}_{0,AA}, \hat{p}_{0,Aa}, \hat{p}_{0,aa}$ *denote the sub-model MLEs, the generalized likelihood ratio is*

$$\Lambda = \left(\frac{\hat{p}_{0,AA}}{\hat{p}_{AA}}\right)^{N_{AA}} \left(\frac{\hat{p}_{0,Aa}}{\hat{p}_{Aa}}\right)^{N_{Aa}} \left(\frac{\hat{p}_{0,aa}}{\hat{p}_{aa}}\right)^{N_{aa}},$$

so

[9]Read more: https://www.nature.com/scitable/knowledge/library/the-hardy-weinberg-principle-13235724/

$$-2\log\Lambda = 2N_{AA}\log\frac{\hat{p}_{AA}}{\hat{p}_{0,AA}} + 2N_{Aa}\log\frac{\hat{p}_{Aa}}{\hat{p}_{0,Aa}} + 2N_{aa}\log\frac{\hat{p}_{aa}}{\hat{p}_{0,aa}}. \tag{1}$$

The full-model MLEs are given by $\hat{p}_{AA} = N_{AA}/n, \hat{p}_{Aa} = N_{Aa}/n$, *and* $\hat{p}_{aa} = N_{aa}/n$, *by Example 1.3.4 from Chapter 2. To find the sub-model MLEs, note that under* H_0, *the multinomial likelihood as a function of* θ *is*

$$\begin{aligned}\text{lik}(\theta) &= \binom{n}{N_{AA}, N_{Aa}, N_{aa}} \left(\theta^2\right)^{N_{AA}} \left(2\theta(1-\theta)\right)^{N_{Aa}} \left((1-\theta)^2\right)^{N_{aa}} \\ &= \binom{n}{N_{AA}, N_{Aa}, N_{aa}} 2^{N_{Aa}}\theta^{2N_{AA}+N_{Aa}}(1-\theta)^{N_{Aa}+2N_{aa}}.\end{aligned}$$

Maximizing the likelihood over parameters (p_{AA}, p_{Aa}, p_{aa}) *belonging to the sub-model is equivalent to maximizing the above over* θ. *Differentiating the logarithm of the above likelihood and setting it equal to 0, we obtain the MLE*

$$\hat{\theta} = \frac{2N_{AA} + N_{Aa}}{2N_{AA} + 2N_{Aa} + 2N_{aa}} = \frac{2N_{AA} + N_{Aa}}{2n}$$

for θ, *which yields the sub-model MLEs*

$$\begin{aligned}\hat{p}_{0,AA} &= \left(\frac{2N_{AA} + N_{Aa}}{2n}\right)^2 \\ \hat{p}_{0,Aa} &= 2\left(\frac{2N_{AA} + N_{Aa}}{2n}\right)\left(\frac{N_{Aa} + 2N_{aa}}{2n}\right) \\ \hat{p}_{0,aa} &= \left(\frac{N_{Aa} + 2N_{aa}}{2n}\right)^2.\end{aligned}$$

Substituting these expressions into equation (1) yields the formula for $-2\log\Lambda$ *in terms of the observed counts* N_{AA}, N_{Aa}, N_{aa}. *The difference in dimensionality of the two models is* $2 - 1 = 1$, *so an approximate level-*α *test would reject* H_0 *when* $-2\log\Lambda$ *exceeds* $\chi^2_1(\alpha)$.

Rice[10] provides an example (Example 8.5.1A) of genotype data from a population of $n = 1029$ individuals in Hong Kong, in which the alleles determine the presence of an antigen in the red blood cell. In this example, $N_{AA} = 342, N_{Aa} = 500, N_{aa} = 187$, and we may calculate $-2\log\Lambda = 0.0325$. Letting F denote the χ^2_1 CDF, the p-value of our test is $1 - F(0.0325) = 0.86$, so there is no significant evidence of deviation from the Hardy-Weinberg equilibrium.

[10]Rice, John A. Mathematical statistics and data analysis. Cengage Learning, 2006.

6.3 Testing in contingency tables

6.3.1 Test of independence

We introduced the generalized likelihood ratio test, and we applied it to an example of testing the hypothesis of Hardy-Weinberg equilibrium in a population at a single diallelic locus. This was an example of testing whether the parameters of a multinomial model satisfy certain additional constraints.

Here is a second example of this type of hypothesis testing problem:

Example 6.3.1. *(Independence test). The following table (from the GSS 2008*[11]*) cross-classifies a random sample of 1972 people by gender and by political party identification:*

	dem	indep	repub
female	422	381	273
male	299	365	232

In this sample, approximately 39% *of females identified as democrat and* 25% *identified as republican, while approximately* 33% *of males identified as democrat and* 26% *identified as republican. Is this significant evidence of an association between gender and party identification in the population from which this sample was drawn?*

Denote the observed counts by N_{ij} *for* $i = 1, 2$ *and* $j = 1, 2, 3$*. We may model these counts as multinomial with* $n = 1972$ *total observations and with outcome probabilities* p_{ij} *for* $i = 1, 2$ *and* $j = 1, 2, 3$*. Denote by* $p_{i\cdot} = \sum_j p_{ij}$ *and* $p_{\cdot j} = \sum_i p_{ij}$ *the marginal row and column probabilities. If there is no association between gender and party identification, then* $p_{ij} = p_{i\cdot}p_{\cdot j}$*. Hence we wish to test the* ***independence null hypothesis***

$$H_0 : p_{ij} = p_{i\cdot} \cdot p_{\cdot j} \text{ for all } i = 1, 2 \text{ and } j = 1, 2, 3.$$

The dimension of this sub-model may be determined as follows: The five row and column marginal probabilities $p_{1\cdot}, p_{2\cdot}, p_{\cdot 1}, p_{\cdot 2}, p_{\cdot 3}$ *specify all of the multinomial cell probabilities under* H_0*. However, they satisfy the constraints* $p_{1\cdot} + p_{2\cdot} = 1$ *and* $p_{\cdot 1} + p_{\cdot 2} + p_{\cdot 3} = 1$*, so this sub-model has dimension* $5 - 2 = 3$*. The full multinomial model has dimension 5, so the generalized likelihood ratio statistic has approximate null distribution* χ^2_2 *(since* $5 - 3 = 2$*).*

To derive the form of the likelihood ratio statistic in the above example, suppose more generally that we observe $(N_1, \ldots, N_k) \sim \text{Multinomial}\,(n, (p_1, \ldots, p_k))$*, and we wish to*

[11] https://gss.norc.org/

test the null hypothesis $H_0 : (p_1, \ldots, p_k) \in \Omega_0$, where Ω_0 represents some sub-model. The multinomial likelihood is given by

$$\text{lik}\,(p_1, \ldots, p_k) = \binom{n}{N_1, \ldots, N_k} \prod_{i=1}^{k} p_i^{N_i}.$$

Letting $\hat{p}_{0,i}$ denote the MLEs in this sub-model Ω_0 and $\hat{p}_i$ denote the MLEs in the full multinomial model, the generalized likelihood ratio is

$$\Lambda = \frac{\text{lik}\,(\hat{p}_{0,1}, \ldots, \hat{p}_{0,k})}{\text{lik}\,(\hat{p}_1, \ldots, \hat{p}_k)} = \prod_{i=1}^{k} \left(\frac{\hat{p}_{0,i}}{\hat{p}_i} \right)^{N_i},$$

so

$$-2 \log \Lambda = 2 \sum_{i=1}^{k} N_i \log \frac{\hat{p}_i}{\hat{p}_{0,i}}.$$

Recall that the full model MLEs are given by $\hat{p}_i = N_i/n$, by Example 1.3.4 of Chapter 2. Let us write $E_i = \hat{p}_{0,i} n$, which denotes the "expected count" for outcome i corresponding to the sub-model MLE $\hat{p}_{0,i}$. Then we obtain the simple formula

$$-2 \log \Lambda = 2 \sum_{i=1}^{k} N_i \log \frac{N_i}{E_i}. \tag{2}$$

Example 6.3.2. *(Independence test (cont'd)). Applying equation (2) to Example 6.3.1, we must compute the sub-model MLEs. Under H_0, the likelihood as a function of the row and column marginal probabilities is*

$$\begin{aligned} \text{lik}\,(p_{1\cdot}, p_{2\cdot}, p_{\cdot 1}, p_{\cdot 2}, p_{\cdot 3}) &= \binom{n}{N_{11}, \ldots, N_{23}} \prod_{i=1}^{2} \prod_{j=1}^{3} (p_{i\cdot} \cdot p_{\cdot j})^{N_{ij}} \\ &= \binom{n}{N_{11}, \ldots, N_{23}} \prod_{i=1}^{2} p_i^{N_{i\cdot}} \prod_{j=1}^{3} p_{\cdot j}^{N_{\cdot j}}, \end{aligned}$$

where $N_{i\cdot} = \sum_j N_{ij}$ and $N_{\cdot j} = \sum_i N_{ij}$ are the row and column marginal counts. Taking the logarithm and introducing Lagrange multipliers for the constraints, we wish to maximize

$$\log \binom{n}{N_{11}, \ldots, N_{23}} + \sum_{i=1}^{2} N_{i\cdot} \log p_{i\cdot} + \sum_{j=1}^{3} N_{\cdot j} \log p_{\cdot j} + \lambda \left(\sum_{i=1}^{2} p_{i\cdot} - 1 \right) + \mu \left(\sum_{j=1}^{3} p_{\cdot j} - 1 \right).$$

Setting the derivatives with respect to $p_{i\cdot}$ and $p_{\cdot j}$ equal to 0 yields the equations $N_{i\cdot}/p_{i\cdot} + \lambda = 0$ and $N_{\cdot j}/p_{\cdot j} + \mu = 0$, so $p_{i\cdot} = -N_{i\cdot}/\lambda$ and $p_{\cdot j} = -N_{\cdot j}/\mu$. Picking the Lagrange multipliers $\lambda = -n$ and $\mu = -n$ enforces the constraints, and we obtain the MLEs $\hat{p}_{i\cdot} = N_{i\cdot/n}/n$ and $\hat{p}_{\cdot j} = N_{\cdot j}/n$. Then the sub-model MLEs for p_{ij} are given by $\hat{p}_{0,ij} = \left(N_{i\cdot/n}/n\right)(N_{\cdot j}/n)$.

For the data of Example 6.3.1, the row and column marginal counts are given by $N_{1\cdot} = 1076$, $N_{2\cdot} = 896$, $N_{\cdot 1} = 721$, $N_{\cdot 2} = 746$, and $N_{\cdot 3} = 505$. Computing the sub-model MLEs $\hat{p}_{0,ij}$ and multiplying by n, we obtain the table of expected counts E_{ij}:

	dem	indep	repub
female	393.4	407.0	275.5
male	327.6	339.0	229.5

Applying equation (2) with the 6 observed and expected counts yields $-2\log\Lambda = 8.31$. Letting F denote the CDF of the χ^2_2 distribution, we obtain a p-value for the generalized likelihood ratio test of $1 - F(8.31) = 0.016$, so there is reasonably strong evidence of an association between gender and party identification.

6.3.2 Test of homogeneity

Consider now a slightly different problem: We have independent count observations from 2 multinomial distributions, each with k outcomes: $(N_1, \ldots, N_k) \sim \text{Multinomial}\,(n, (p_1, \ldots$ and $(M_1, \ldots, M_k) \sim \text{Multinomial}\,(m, (q_1, \ldots, q_k))$, where n and m are known sample sizes. We wish to test the **homogeneity null hypothesis**

$$H_0 : p_i = q_i \text{ for all } i = 1, \ldots, k.$$

Example 6.3.3. *(Homogeneity test). This example is from Rice[12] Section 13.3. When Jane Austen died, she left the novel Sandition partially completed. An admirer finished the novel, attempting to emulate Jane Austen's style. The following table counts the occurrences of six different short words in Chapters 1 and 6 of Sandition, written by Austen, and in Chapters 12 and 24 of Sandition, written by the admirer:*

	a	an	this	that	with	without
Ch. 1 and 6	101	11	15	37	28	10
Ch. 12 and 24	83	29	15	22	43	4

Is there a significant difference between the relative frequencies of these words between the two authors?

[12]Rice, John A. Mathematical statistics and data analysis. Cengage Learning, 2006.

Let us model the counts from Chapters 1 and 6 as $\text{Multinomial}\,(202, (p_1, \ldots, p_6))$ *and those from Chapters 12 and 24 as Multinomial* $(196, (q_1, \ldots, q_6))$*. Then we wish to test the homogeneity null hypothesis that* $p_i = q_i$ *for all* $i = 1, \ldots, 6$*.*

To derive the generalized likelihood ratio test, note that the joint likelihood of all parameters is the product of the two multinomial likelihoods:

$$\text{lik}\,(p_1, \ldots, p_k, q_1, \ldots, q_k) = \binom{n}{N_1, \ldots, N_k} \prod_{i=1}^{k} p_i^{N_i} \times \binom{m}{M_1, \ldots, M_k} \prod_{i=1}^{k} q_i^{M_i}. \quad (3)$$

Let $\hat{p}_i$ *and* $\hat{q}_i$ *denote the full model MLEs, and let* $\hat{p}_{0,i} = \hat{q}_{0,i}$ *denote the sub-model MLEs. Then the generalized likelihood ratio statistic is*

$$\Lambda = \prod_{i=1}^{k} \left(\frac{\hat{p}_{0,i}}{\hat{p}_i} \right)^{N_i} \prod_{i=1}^{k} \left(\frac{\hat{q}_{0,i}}{\hat{q}_i} \right)^{M_i},$$

so

$$-2 \log \Lambda = 2 \sum_{i=1}^{k} \left(N_i \log \frac{\hat{p}_i}{\hat{p}_{0,i}} + M_i \log \frac{\hat{q}_i}{\hat{q}_{0,i}} \right). \quad (4)$$

In the full model with two independent and unconstrained multinomial distributions, the MLEs are simply $\hat{p}_i = N_i/n$ *and* $\hat{q}_i = M_i/m$*. Letting* $E_i = \hat{p}_{0,i} n$ *and* $F_i = \hat{q}_{0,i} m$ *denote the expected counts in the sub-model, we may write the above in the simple form*

$$-2 \log \Lambda = 2 \sum_{i=1}^{k} \left(N_i \log \frac{N_i}{E_i} + M_i \log \frac{M_i}{F_i} \right).$$

To compute the above statistic, we need to compute the sub-model MLEs $\hat{p}_{0,i} = \hat{q}_{0,i}$*. Under* H_0*, the likelihood in equation (3) simplifies to*

$$\text{lik}\,(p_1, \ldots, p_k) = \binom{n}{N_1, \ldots, N_k} \binom{m}{M_1, \ldots, M_k} \prod_{i=1}^{k} p_i^{N_i + M_i}.$$

Taking the logarithm and introducing a Lagrange multiplier for the constraint $p_1 + \ldots + p_k = 1$ *, we wish to maximize*

$$\log \left(\binom{n}{N_1, \ldots, N_k} \binom{m}{M_1, \ldots, M_k} \right) + \sum_{i=1}^{k} (N_i + M_i) \log p_i + \lambda \left(\sum_{i=1}^{k} p_i - 1 \right).$$

Setting the derivatives with respect to p_i equal to 0 , we obtain the equations $(N_i + M_i)/p_i + \lambda = 0$, so $p_i = -(N_i + M_i)/\lambda$. Choosing the Lagrange multiplier $\lambda = -(n+m)$ enforces the constraints, and we obtain the sub-model MLEs $\hat{p}_{0,i} = \hat{q}_{0,i} = (N_i + M_i)/(n + m)$.

Example 6.3.4. *(Homogeneity test (cont'd)). In the data of Example 6.3.3, we have the marginal word counts $N_1 + M_1 = 184, N_2 + M_2 = 40, N_3 + M_3 = 30, N_4 + M_4 = 59$, $N_5 + M_5 = 71$, and $N_6 + M_6 = 14$. This yields the table of expected counts*

	a	an	this	that	with	without
Ch. 1 and 6	93.4	20.3	15.2	29.9	36.0	7.1
Ch. 12 and 24	90.6	19.7	14.8	29.1	35.0	6.9

Applying equation (4) with the observed and expected counts, we obtain $-2\log\Lambda = 19.8$. The dimensionality of the sub-model in this example is 5 (6 parameters minus 1 constraint), and the dimensionality of the full model is 10 (12 parameters minus 2 constraints), so the null distribution of $-2\log\Lambda$ is approximately χ^2_5. Letting F denote the CDF of the χ^2_5 distribution, we obtain a p-value of $1 - F(19.8) = 0.0014$, so there is significant evidence of a difference in writing style between Austen and her admirer.

Remark 6.1.

- *In both Examples 6.3.1 and 6.3.3, we wanted to test whether there is a significant difference in the relative frequencies between the two rows. The distinction between these examples is only in the sampling design/modeling assumption: In Example 6.3.1, we treated the counts from all rows as observations from a single multinomial distribution, because (we believe that) the GSS 2008 survey sampled a fixed total number of people rather than a fixed number of people of each gender. In Example 6.3.3, we modeled each row as a separate multinomial distribution with a fixed row sum.*

- *In fact, the table of expected counts, generalized likelihood ratio statistic, and degrees of freedom for the test are all the same under the two different modeling assumptions (although we derived them in two different ways), so the tests of independence and of homogeneity are procedurally the same, and the distinction between these is sometimes blurred in practice.*

7 Experimental design

7.1 Steps of a statistical study

A "typical" statistical study might consist of the following steps:

1. Identify/formulate the question of interest
2. Design an experiment or study to collect data that addresses this question
3. Clean, visualize and explore the data
4. Draw an inference from the data to answer the original question

So far, our focus has been on Step 4. (We discussed briefly ideas such as hanging histogram plots and QQ plots for Step 3.) Now we'll discuss some aspects of Step 2, in the context of two-sample hypothesis testing. We try to address the following questions:

- How can we eliminate or minimize the influence of confounding factors?[13]
- How can we reason about the size of the study needed to identify an effect of interest?
- How can we design the experiment so as to maximize the chance of identifying this effect?

7.2 Case study: Peer grading students in statistics course

- **Context**: Grading student homework assignments in large classes is time-consuming and costly, perhaps prohibitively so in Massive Open Online Courses (MOOCs) with thousands or tens of thousands of students.
- **Possible solution**: Have students grade each other (peer grading).
- **Question of interest**: Can peer grading actually increase student learning?

Justice Anthony Kennedy, in Supreme Court case *Owasso v. Valvo*: "Correcting a classmate's work can be as much a part of the assignment as taking the test itself. It is a way to teach material again in a new context, and it helps show students how to assist and respect fellow pupils."

[13] see examples in Rice, John A. Mathematical statistics and data analysis. Cengage Learning, 2006., Section 11.4.

7.2.1 A simple design

Suppose there are 300 students undertaking a statistics course. Divide them into two groups, "peer-grading" and "control". Have only the students in the peer-grading group grade their peers, and compare learning (e.g. test scores) between the two groups at the end of the quarter.

Problem: Student performance is influenced by many confounding factors – their class year, previous coursework and knowledge of statistics, etc.

Simple solution: Randomly assign students to the two groups, so that confounding factors tend to be balanced between the groups. For this design, we might use a two-sample t-test:

Let $X_1, \ldots, X_n$ be final exam scores of the peer-grading group, $Y_1, \ldots, Y_m$ those of the control group. Supposing that $X_1, \ldots, X_n \sim \mathcal{N}\left(\mu_X, \sigma^2\right), Y_1, \ldots, Y_m \sim \mathcal{N}\left(\mu_Y, \sigma^2\right)$, test

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X > \mu_Y$$

using the two-sample T-statistic $T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$, where S_p^2 is the pooled sample variance discussed in Section 4.1.

What is the chance that we identify a significant effect (reject H_0)?

Calculating the power

Here $n+m = 300$ is fairly large, so we expect S_p^2 to be a very accurate estimate of σ^2. Let's assume for simplicity that we know σ^2, and perform the test using

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}.$$

Recall,

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right).$$

Under H_0, $Z \sim \mathcal{N}(0,1)$, so a one-sided test rejects when $Z > z(\alpha)$.

Under H_1, $Z \sim \mathcal{N}(d,1)$, where

$$d = \frac{\mu_X - \mu_Y}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}.$$

The power of the test increases with d:

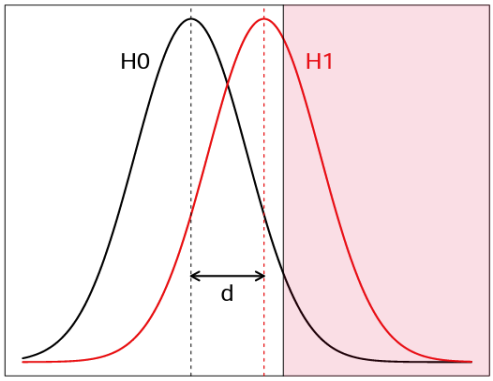


Figure 3: Distributions of Z under H_0 and H_1.

The separation

$$d = \frac{\mu_X - \mu_Y}{\sigma}\sqrt{\frac{1}{\frac{1}{n}+\frac{1}{m}}}$$

is determined by:

- The real difference in mean test scores $\mu_X - \mu_Y$.
- The standard deviation of test scores (i.e., "noise" level) σ.
- The sample sizes n and m.

The quantity $\frac{\mu_X - \mu_Y}{\sigma}$ is called the **effect size** – it measures the size of the mean difference in terms of the number of standard deviations of the noise.

The power of the test is

$$\mathbb{P}_{H_1}[Z > z(\alpha)] = \mathbb{P}_{H_1}[Z - d > z(\alpha) - d] = 1 - \Phi(z(\alpha) - d),$$

where $\Phi(x)$ is the standard normal CDF, and we used the fact that $Z - d \sim \mathcal{N}(0,1)$ under H_1.

Subject to the constraint of $n+m = 300$ total students, d is maximized when we choose $n = m = 150$ students per group. The effect size identified by the study (in retrospect) was 0.11. So

$$d = \frac{\mu_X - \mu_Y}{\sigma}\sqrt{\frac{n}{2}} = 0.95.$$

At level $\alpha = 0.05$, the above power is only 0.244! In other words, had we done this experiment, we would have only had a 24% chance of rejecting H_0 at level $\alpha = 0.05$.

Typical p-value

We can also think in terms of the p-value we would have obtained. If the test statistic we observed were Z, then the p-value would be the upper tail probability

$$P = 1 - \Phi(Z).$$

(P and Z here are both random, depending on the outcome of the experiment.)

Under H_1, $Z \sim \mathcal{N}(d, 1)$, so the median value of Z is d. Since $x \mapsto 1 - \Phi(x)$ is monotone (decreasing), the median value of P is $1 - \Phi(d)$. So a "typical" p-value from this experiment would have been $1 - \Phi(d)$. For $d = 0.95$, this p-value is 0.17.

Both of these calculations indicate that the study would be **underpowered** – the effect size is too small to be detected with statistical significance if the sample size is 300 students.

How many samples are needed?

Suppose we would like the power to be much larger, say 0.9, under a level $\alpha = 0.05$ test. How many students would we need? If we have n students in each of the peer-grading and control groups, set

$$0.9 = 1 - \Phi(z(\alpha) - d) = 1 - \Phi\left(z(0.05) - 0.11\sqrt{\frac{n}{2}}\right)$$

and solve for n:

$$\Phi\left(z(0.05) - 0.11\sqrt{\frac{n}{2}}\right) = 0.1$$

$$\Rightarrow z(0.05) - 0.11\sqrt{\frac{n}{2}} = \Phi^{-1}(0.1) = -z(0.1)$$

$$\Rightarrow n = 2\left(\frac{z(0.05) + z_{0.1}}{0.11}\right)^2 \approx 1416$$

We would need $2n \approx 2832$ total students. This amounts to doing this experiment for $5 - 10$ years of students from the statistics course.

Remark 7.1. ***Effect sizes in education***

The previous calculations assumed we knew the effect size was 0.11. In reality, we don't know this ahead of time. However, we can compare to what we know:

- *Classroom discussion – 0.82*

- *Computer-assisted instruction – 0.45*
- *Teacher education – 0.12*
- *Charter schools – 0.07*

These numbers are from the 2015 Hattie ranking[14]*, which lists effect sizes for 195 different educational influences/approaches, determined from aggregating previous experimental studies. In education, an effect size larger than* 0.4 *is typically considered strong.*

7.2.2 A different design to improve power

Main problem: There is too much variation in student performance, compared to the size of the improvement from peer-grading.

Idea: Compare each student to himself/herself.

Implementation: Divide statistics course students into 2 units [15], with a quiz at the end of each unit. Assign each student to do peer-grading for one unit, and no peer-grading for the other unit. (To handle the possible confounding factor that one exam is easier than the other, randomly choose which unit each student does peer-grading.) In other words, set up an experiment with paired samples rather than two independent samples.

Calculating the power for paired samples: Why does this help (and how much does it help by)?

Suppose there are n students. Let $X_1, \ldots, X_n$ be their quiz scores in the peer-grading unit, and $Y_1, \ldots, Y_n$ their scores in the control unit. For this design, we might use a one-sample (a.k.a. paired two-sample) t-test:
Let $D_i = X_i - Y_i$, and reject H_0 for large values of the t-statistic

$$T = \frac{\sqrt{n}\bar{D}}{S}.$$

Here, $\bar{D}$ and S^2 are the sample mean and variance of the D_i's.

Assume $X_i \sim \mathcal{N}\left(\mu_X, \sigma^2\right)$ and $Y_i \sim \mathcal{N}\left(\mu_Y, \sigma^2\right)$, as before. Since X_i and Y_i correspond to the same student, they are likely very correlated. Let's suppose (X_i, Y_i) is bivariate

[14] https://visible-learning.org/hattie-ranking-influences-effect-sizes-learning-achievement/
[15] The real study used 4 units instead of 2.

normal with correlation ρ:

$$(X_i, Y_i) \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}\right).$$

Then $D_i = X_i - Y_i$ is normally distributed, with mean $\mathbb{E}[D_i] = \mu_X - \mu_Y$ and variance

$$\begin{aligned}
\mathrm{Var}[D_i] &= \mathrm{Cov}[X_i - Y_i, X_i - Y_i] \\
&= \mathrm{Cov}[X_i, X_i] - \mathrm{Cov}[X_i, Y_i] - \mathrm{Cov}[Y_i, X_i] + \mathrm{Cov}[Y_i, Y_i] \\
&= \sigma^2 - \rho\sigma^2 - \rho\sigma^2 + \sigma^2 \\
&= 2\sigma^2(1-\rho).
\end{aligned}$$

Since n is large, S^2 should be very close to $\mathrm{Var}[D_i] = 2\sigma^2(1-\rho)$. Let's suppose again for simplicity that we know $2\sigma^2(1-\rho)$, and consider the test statistic

$$Z = \frac{\sqrt{n}\bar{D}}{\sqrt{2\sigma^2(1-\rho)}}.$$

We have $\bar{D} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{2\sigma^2(1-\rho)}{n}\right)$.

Under H_0, $Z \sim \mathcal{N}(0,1)$, so a level-α test rejects when $Z > z(\alpha)$.

Under H_1, $Z \sim \mathcal{N}(d,1)$, where

$$d = \frac{\mu_X - \mu_Y}{\sigma}\sqrt{\frac{n}{2(1-\rho)}}.$$

Compared to having two independent samples of size n (one peer-grading, one control), we gain a factor of $1/\sqrt{1-\rho}$ in d. You can think of this as either reducing the effective variance from σ^2 (in the case of unpaired samples) to $\sigma^2(1-\rho)$ (in the case of paired samples), or as increasing the effective sample size from n (in the case of unpaired samples) to $n/(1-\rho)$ (in the case of paired samples). The factor $1-\rho$ is called the **relative efficiency** of the unpaired design to the paired design.

Example 7.2.1. *If $\rho = 0.9$, then the relative efficiency is 0.1, and a paired design with n pairs yields the same power as an unpaired design with two independent samples of size $10n$.*

Examples of paired designs

- Before-and-after studies on the same subjects

- Twin studies
- Subject matching by covariates (e.g., in a medical study, matching by age, weight, severity of condition, etc.)

Matching by covariates was also used in the statistics students experiment: Rather than randomly choosing, for each student, which unit they did peer-grading, each student was paired with the "most similar" other student based on gender, race, previous statistics background, class year, etc. using a matching algorithm. One student in each pair was then randomly assigned to peer grade in unit 1, and the other to peer grade in unit 2. Pairing by covariates is a special case of a **randomized block design**, which groups subjects into blocks having similar characteristics.

Summary of the study

- The estimated (short-term) effect size was 0.11. Despite the small size of the effect, it was found to be statistically significant with p-value 0.002.
- Long-term effect was assessed by comparing performance on the questions corresponding to each unit in the final exam; the estimated effect size was 0.12, with p-value 0.001.

Conclusion: Peer grading yielded a small but real improvement in student learning[16].

7.3 Addendum: S^2 is close to σ^2 for large n

As $n \to \infty$, the sample variance $S^2 \to \sigma^2$ in probability. Why?

Suppose $X_1, \ldots, X_n$ are IID with mean 0 and variance σ^2.

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right) \\ &= \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \bar{X}^2. \end{aligned}$$

As $n \to \infty$, $\frac{n}{n-1} \to 1$. Also by the LLN, $\frac{1}{n} \sum_{i=1}^{n} X_i^2 \to \sigma^2$ and $\bar{X} \to 0$ in probability.

[16]Details available at: Sun, D. L., Harris, N., Walther, G., & Baiocchi, M. (2015). "Peer assessment enhances student learning: The results of a matched randomized crossover experiment in a college statistics class." PloS one.

The functions $(x, y) \mapsto x - y$ and $(x, y) \mapsto xy$ are continuous. So if $X_n \to a$ and $Y_n \to b$ in probability, then the Continuous Mapping Theorem implies $X_n - Y_n \to a - b$ and $X_n Y_n \to ab$ in probability. Then

$$S^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \bar{X}^2 \to 1 \cdot \sigma^2 - 1 \cdot 0 \cdot 0 = \sigma^2$$

in probability. Clearly this also holds if $X_1, \ldots, X_n$ are IID with mean μ and variance σ^2, because S^2 doesn't depend on μ. Note that we didn't assume that the X_i's are normally distributed – this argument holds as long as $X_1, \ldots, X_n$ are IID with finite variance.

8 Testing multiple hypotheses

8.1 The multiple testing problem: More p-values, more problems

XKCD by Randall Monroe[17] takes esoteric math and science concepts and turns them into jokes. In one example, Monroe tackles the issue of multiple hypothesis testing: If you test many hypotheses simultaneously without adjusting your significance cutoff (e.g., $p < 0.05$), false positives are going to happen more than you might expect.

Motivating example: In the related edition of XKCD, two characters want to know if jelly beans cause acne. Scientists investigate this claim and find no link between jelly beans and acne. That is, the scientists test the null hypothesis, "There is no statistically significant relationship between jelly bean consumption and acne." The results will not surprise you.

Objective of the case study: In this joke example, the scientists test one hypothesis, calculating one p-value and comparing that one p-value to a critical value (here 0.05). No problem so far. But, what if I am concerned that one specific color out of a possible, say, 20 jelly bean colors cause acne?

Findings of the case study: So green jelly beans cause acne? We can see the headlines now. Notice that in part of this joke front-page news there is a comment "only 5% chance of coincidence." Is that right? If the scientists had tested a single hypothesis, then yes. However, that's not what happened. The scientists tested 20 hypotheses. So what are the odds this result happened by chance?
Thoughts on the case study: p-value tells you the likelihood of getting a result as extreme or more extreme by chance. That means for a single hypothesis test, the

[17] https://xkcd.com/882/


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

p-value tells you the likelihood of getting something like your result by chance. What about if we tested 20 independent hypotheses with a cutoff of 0.05? In that case,

$$P(\text{at least one false positive}) = 1 - P(\text{no results are significant}) = 1 - (1 - 0.05)^{20} \approx 0.64.$$

There is a 64% chance of at least one false positive. Said another way, it is more likely than not that this experiment will yield at least one false positive just by chance. What is the possible remedy?

The simplest approach is to divide your cut-off value by the number of simultaneous hypotheses. This process is called a Bonferroni correction In this case, that would be

$$\text{Cutoff} = \frac{0.05}{20} = 0.0025$$

You may think, "That's a very strict cutoff." You're right. This cutoff will do a great job of preventing false positives. In fact, we can prove it.

$$P(\text{ at least one false positive }) = 1 - P(\text{ no results are significant }) = 1 - (1 - 0.0025)^{20} \approx 0.$$

This number, 0.0488, can be thought of as the cut-off equivalent. If we were to somehow condense all 20 tests into 1, the cutoff for this test would be 0.488. However, as you might expect, this process results in more false negatives than would be expected from a single hypothesis test. In fact, you can prove that the false negative rate tends toward 1 as the number of tests increases[18]. That is, if you do a lot of simultaneous tests with this method, you'll fail to reject the null hypothesis nearly every time, regardless of whether there is actually a relationship in our data.

The Bonferoni correction is still useful, though. If having even one false positive would mean disaster for your work, then the Bonferoni correction may be the way to go, as it is quite conservative. Likewise, if you are testing only a small number of hypotheses (say < 25) and we expect (based on prior knowledge) that only one or two are true, the Bonferoni correction may be the way to go. One option is to control the False Discovery Rate (FDR).[19] For a theoretical description, see Benjamini and Hochberg (1995).[20]

Multiple testing problem: If we test n true null hypotheses at level α, then on average we'll still (falsely) reject αn of them. Examples are as follows:

[18]Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. Journal of the American Statistical Association, 99(465), 96-104.

[19]https://www.biostathandbook.com/multiplecomparisons.html

[20]Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289-300.

- Test the safety of a drug in terms of a dozen different side effects.
- Test whether a disease is related to 10,000 different gene expressions.

What are some ways we can think about acceptance/rejection errors across multiple hypothesis tests/experiments? What statistical procedures can control these measures of errors?

8.2 The Bonferroni correction

Consider testing n different null hypotheses $H_0^{(1)}, \ldots, H_0^{(n)}$, all of which are, in fact, true. One goal we might set is to ensure

$$\mathbb{P}\left[\text{reject any null hypothesis}\right] \leq \alpha.$$

A simple and commonly-used method of achieving this is called the **Bonferroni** method: Perform each test at significance level α/n, instead of level α. Verification:

$$\begin{aligned}
\mathbb{P}\left[\text{Reject any null hypothesis}\right] &= \mathbb{P}\left[\left\{\text{ Reject } H_0^{(1)}\right\} \cup \ldots \cup \left\{\text{ Reject } H_0^{(n)}\right\}\right] \\
&\leq \mathbb{P}\left[\text{ Reject } H_0^{(1)}\right] + \ldots + \mathbb{P}\left[\text{ Reject } H_0^{(n)}\right] \\
&= \frac{\alpha}{n} + \ldots + \frac{\alpha}{n} = \alpha
\end{aligned}$$

8.3 Family-wise error rate

More generally, suppose we test n null hypotheses, n_0 of which are true and $n - n_0$ of which are false. The results of the tests might be tabulated as follows:

	H_0 is true	H_0 is false	Total
Reject H_0	V	S	R
Accept H_0	U	T	$n - R$
Total	n_0	$n - n_0$	n

$R = \#$ rejected null hypotheses,
$V = \#$ type I errors,
$T = \#$ type II errors.

Remark 8.1. *We consider n_0 and $n - n_0$ to be fixed quantities. The number of hypotheses we reject, R, as well as the cell counts U, V, S, T, are random, as they depend on the data observed in each experiment.*

The **family-wise error rate** (FWER) is the probability of falsely rejecting at least one true null hypothesis,

$$\mathbb{P}[V \geq 1].$$

A procedure controls FWER at level α if $\mathbb{P}[V \geq 1] \leq \alpha$, regardless of the (possibly unknown) number of true null hypotheses n_0.

Bonferroni controls FWER: Without loss of generality, let $H_0^{(1)}, \ldots, H_0^{(n_0)}$ be the true null hypotheses.

$$\begin{aligned}
\mathbb{P}[V \geq 1] &= \mathbb{P}\left[\left\{ \text{ Reject } H_0^{(1)} \right\} \cup \ldots \cup \left\{ \text{ Reject } H_0^{(n_0)} \right\}\right] \\
&\leq \mathbb{P}\left[\text{ Reject } H_0^{(1)} \right] + \ldots + \mathbb{P}\left[\text{ Reject } H_0^{(n_0)} \right] \\
&= \frac{\alpha}{n} + \ldots + \frac{\alpha}{n} = \frac{\alpha n_0}{n} \leq \alpha.
\end{aligned}$$

8.4 Thinking in terms of p-values

Many multiple-testing procedures are formulated as operating on the p-values returned by individual tests, rather than on the original data or the test statistics that were used.

For example, Bonferroni may be described as follows: Reject those null hypotheses whose corresponding p-values are at most α/n.

Key advantages:

- Abstracts away details about how individual tests were performed.
- Applicable regardless of which tests/test statistics were used for each experiment.
- Allows for meta-analyses of previous experiments without access to the original data.

8.5 The null distribution of a p-value

Suppose a null hypothesis H_0 is true, and we perform a statistical test of H_0 and obtain a p-value P. What is the distribution of P?

If our test statistic T has a continuous distribution under H_0 with CDF F, and we reject for small values of T, then the p-value is just the lower tail probability

$$P = F(T).$$

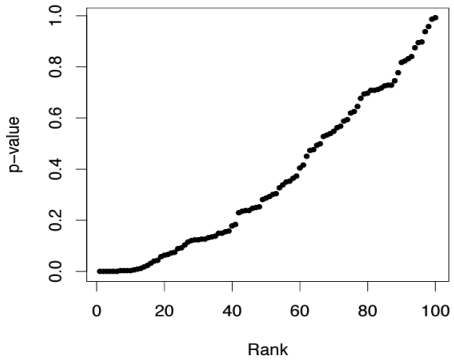
For any $t \in (0,1)$

$$\mathbb{P}[P \leq t] = \mathbb{P}[F(T) \leq t] = \mathbb{P}\left[T \leq F^{-1}(t)\right] = F\left(F^{-1}(t)\right) = t.$$

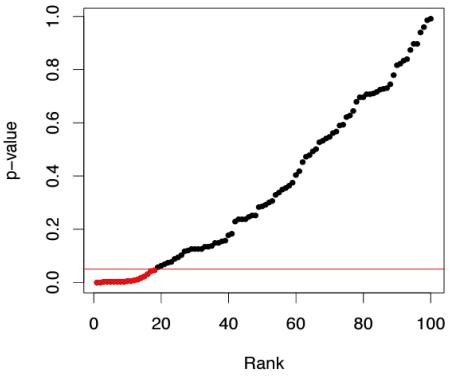
So $P \sim \text{Uniform}(0,1)$. Similarly, $P \sim \text{Uniform}(0,1)$ if we reject for large T, or both large and small T.[21]

8.6 Ordered p-value plots

We can understand multiple testing procedures visually in terms of the plot of the ordered p-values (sorted from smallest to largest):

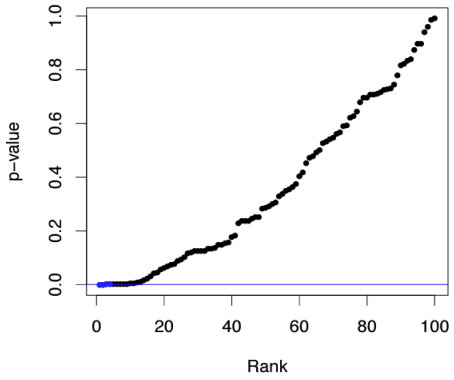


Applying each test at level 0.05, we reject the null hypotheses corresponding to the below 18 red points.



Applying the Bonferroni correction, we reject null hypotheses with p-value less than 0.0005, corresponding to the below 4 blue points.

[21] If T has a discrete distribution under H_0, then so does P, so the null distribution of P wouldn't be exactly Uniform $(0,1)$.



8.7 False discovery rate

	H_0 is true	H_0 is false	Total
Reject H_0	V	S	R
Accept H_0	U	T	$n - R$
Total	n_0	$n - n_0$	n

Controlling the FWER $\mathbb{P}[V \geq 1]$ may be too conservative and greatly reduce our power to detect real effects, especially when n (the total number of tested hypotheses) is large.

In many modern "large-scale testing" applications, the focus has shifted to the **false-discovery proportion** (FDP)

$$\text{FDP} = \begin{cases} \frac{V}{R} & R \geq 1 \\ 0 & R = 0, \end{cases}$$

and on procedures that control its expected value $\mathbb{E}[\text{FDP}]$, called the **false-discovery rate** (FDR).

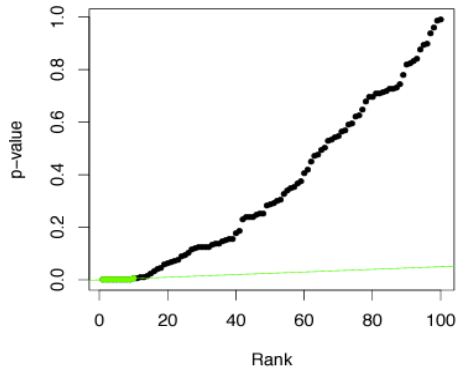
8.8 FWER vs. FDR

Controlling FDR is a shift in paradigm – we are willing to tolerate some type I errors (false discoveries), as long as most of the discoveries we make are still true.

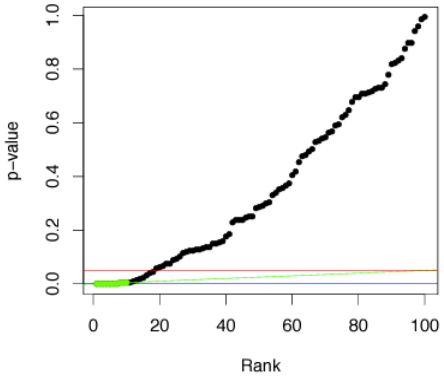
It has been argued that in applications where the statistical test is thought of as providing a "definitive answer" for whether an effect is real, FWER control is still the correct objective. In contrast, for applications where the statistical test identifies candidate effects that are likely to be real and which merit further study, it may be better to target FDR control.

8.9 The Benjamini-Hochberg procedure

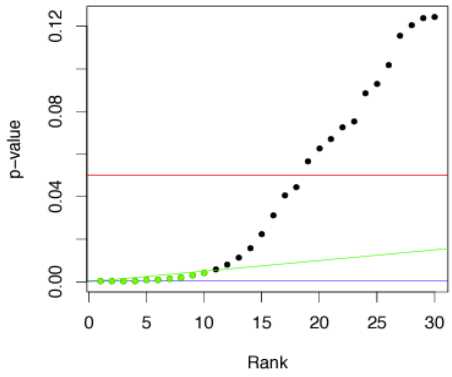
The **Benjamini-Hochberg (BH)** procedure compares the sorted p-values to a diagonal cutoff line, finds the largest p-value that still falls below this line, and rejects the null hypotheses for the p-values up to and including this one.



To control FDR at level q, the diagonal cutoff line is set to equal the Bonferroni level q/n at the smallest p-value and to equal the uncorrected level q at the largest p-value.



Here's the same picture, zoomed in to the 30 smallest p-values. In this example, the BH procedure rejects the 10 null hypotheses corresponding to the points in green.



Formally, the BH procedure at level q is defined as follows:

1. Sort the p-values. Call them $P_{(1)} \leq \ldots \leq P_{(n)}$.

2. Find the largest r such that $P_{(r)} \leq \frac{qr}{n}$.

3. Reject the null hypotheses $H_{(1)}, \ldots, H_{(r)}$.

Theorem 8.1. *(Benjamini and Hochberg (1995)[22]) Consider tests of n null hypotheses, n_0 of which are true. If the test statistics (or equivalently, p-values) of these tests are independent, then the FDR of the above procedure satisfies*[23]

$$\text{FDR} \leq \frac{n_0 q}{n} \leq q.$$

Motivation of the BH procedure: For each $\alpha \in (0,1)$, let $R(\alpha)$ be the number of p-values $\leq \alpha$. If we reject hypotheses with p-value $\leq \alpha$, then we expect (on average) to falsely reject αn_0 null hypotheses, since the null p-values are distributed as Uniform $(0,1)$. So we might estimate the false discovery proportion by

$$\alpha n_0 / R(\alpha)$$

As we don't know n_0, let's take the conservative upper-bound

$$\alpha n / R(\alpha)$$

If we set $\alpha = P_{(r)}$, the r th largest p-value, then $\alpha n / R(\alpha) \leq q$ exactly when $P_{(r)} \leq qr/n$. So the BH procedure chooses α (in a data-dependent way) so as to reject as many hypotheses as possible, subject to the constraint $\alpha n / R(\alpha) \leq q$.

Proof. Let's prove (more formally) the theorem that BH controls the FDR. For any event $\mathcal{E}$, we use the indicator notation

$$\mathbb{1}\{\mathcal{E}\} = \begin{cases} 1 & \mathcal{E} \text{ holds} \\ 0 & \mathcal{E} \text{ does not hold.} \end{cases}$$

Without loss of generality, order the n null hypotheses $H_0^{(1)}, \ldots, H_0^{(n)}$ so that the first n_0 of them are true nulls. Then

$$\begin{aligned} \text{FDR} &= \mathbb{E}[\text{FDP}] \\ &= \mathbb{E}\left[\sum_{r=1}^{n} \frac{V}{r} \mathbb{1}\{R = r\}\right] \\ &= \mathbb{E}\left[\sum_{r=1}^{n} \sum_{j=1}^{n_0} \mathbb{1}\left\{ \text{ reject } H_0^{(j)} \right\} \frac{1}{r} \mathbb{1}\{R = r\}\right], \end{aligned}$$

[22] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: series B (Methodological), 57(1), 289-300.

[23] FDR control is not guaranteed if the test statistics are dependent.

(where we have noted $V = \sum_{j=1}^{n_0} \mathbb{1}\left\{\text{reject } H_0^{(j)}\right\}$).

Applying linearity of expectation,

$$\begin{aligned}\text{FDR} &= \sum_{r=1}^{n}\sum_{j=1}^{n_0}\frac{1}{r}\mathbb{E}\left[\mathbb{1}\left\{\text{ reject } H_0^{(j)}\right\}\mathbb{1}\{R=r\}\right] \\ &= \sum_{r=1}^{n}\sum_{j=1}^{n_0}\frac{1}{r}\mathbb{P}\left[\text{ reject } H_0^{(j)} \text{ and } R=r\right].\end{aligned}$$

For fixed j, let $P^*_{(1)} \leq \ldots \leq P^*_{(n-1)}$ be the sorted $n-1$ p-values other than P_j. Then the BH procedure rejects r total hypotheses including $H_0^{(j)}$ if and only if $P_j \leq \frac{qr}{n}$ and the following event holds:

$$\mathcal{E}^{(r)} := \left\{P^*_{(1)},\ldots,P^*_{(r-1)} \leq \frac{qr}{n},\ P^*_{(r)} > \frac{q(r+1)}{n},\ P^*_{(r+1)} > \frac{q(r+2)}{n},\ldots,\ P^*_{(n-1)} > q\right\}.$$

As the p-values are independent, P_j is independent of $P^*_{(1)},\ldots,P^*_{(n-1)}$. Furthermore, $P_j \sim \text{Uniform}(0,1)$. So

$$\begin{aligned}\text{FDR } &= \sum_{r=1}^{n}\sum_{j=1}^{n_0}\frac{1}{r}\mathbb{P}\left[P_j \leq \frac{qr}{n} \text{ and } \mathcal{E}^{(r)} \text{ holds}\right] \\ &= \sum_{j=1}^{n_0}\sum_{r=1}^{n}\frac{1}{r}\mathbb{P}\left[P_j \leq \frac{qr}{n}\right]\mathbb{P}\left[\mathcal{E}^{(r)} \text{ holds}\right] \\ &= \sum_{j=1}^{n_0}\sum_{r=1}^{n}\frac{1}{r}\frac{qr}{n}\mathbb{P}\left[\mathcal{E}^{(r)} \text{ holds}\right] = \frac{q}{n}\sum_{j=1}^{n_0}\sum_{r=1}^{n}\mathbb{P}\left[\mathcal{E}^{(r)} \text{ holds}\right].\end{aligned}$$

Finally, note that (for any fixed j) the events $\mathcal{E}^{(1)},\ldots,\mathcal{E}^{(n)}$ are mutually exclusive – $\mathcal{E}^{(r)}$ holds if and only if the largest index k such that $P^*_{(k)} \leq \frac{q(k+1)}{n}$ is exactly $k = r-1$ (with $\mathcal{E}^{(1)}$ holding if $P^*_{(k)} > \frac{q(k+1)}{n}$ for all k), and this is true for exactly one value of $r \in \{1,\ldots,n\}$. So

$$\sum_{r=1}^{n}\mathbb{P}\left[\mathcal{E}^{(r)} \text{ holds}\right] = 1.$$

Hence

$$\text{FDR} \leq \frac{q}{n}\sum_{j=1}^{n_0}1 = \frac{qn_0}{n} \leq q.$$

□